

# CS6120/CS4120: Natural Language Processing

Instructor: Prof. Lu Wang

College of Computer and Information Science

Northeastern University

Webpage: [www.ccs.neu.edu/home/luwang](http://www.ccs.neu.edu/home/luwang)

# Logistics

- Arrangement for this week and next week
  - Today: QA cont'd, summarization
  - Oct 30: guest lecture “**Developing NLP measures for social science applications**” by Sarah Shugars (PhD Candidate in Network Science), **attendance required**
  - Nov 4: only one office hour at 2:50-4:30pm, Churchill Hall 101
  - Nov 6: prelim, 2:50-4:30pm, Churchill Hall 101, open book

# Project progress report

1. 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.
  2. Describe data preprocessing process. This includes data cleaning, selection, feature generation or other representation you have used, etc.
  3. What methods or models you have tried towards the project goal? And why do you choose the methods (you can include related work on similar task or relevant tasks)?
  4. What results you have achieved up to now based on your proposed evaluation methods? What worked and what didn't work?
  5. How can you improve your models? What are the next steps?
- Grading: For 2-5, each aspect will take 25 points.
  - 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](http://icml.cc/2015/?page_id=151).
  - Each group only needs to submit one copy.
  - Due Nov 18, 11:59pm
  - Feel free to reach out to the staffs you have any question!

# 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



News articles



Emails



Social Media Streams



Scientific Articles



Books



Websites

# Speech Summarization

Phone Conversation



Lecture



Meeting



Talk Shows



Broadcast News



Chat



Classroom



Radio News

- “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]

# Example: “what is keto diet”

## The Ketogenic Diet: A Detailed Beginner's Guide to Keto - Healthline

<https://www.healthline.com/nutrition/ketogenic-diet-101> ▼

Jul 30, 2018 - The ketogenic diet is a very **low-carb**, high-fat diet that shares many similarities with the **Atkins** and **low-carb** diets. It involves drastically reducing carbohydrate intake and replacing it with fat. This reduction in **carbs** puts your body into a metabolic state called ketosis.

[What It Is](#) · [Types](#) · [Other Benefits](#) · [Sample Meal Plan](#)

## 16 Foods to Eat on a Ketogenic Diet - Healthline

<https://www.healthline.com/nutrition/ketogenic-diet-foods> ▼

Jan 23, 2017 - A **ketogenic diet** is a very low-carb diet with numerous health benefits. Here are 16 healthy and nutritious foods you can eat on this diet.

## 8 Steps Beginners Should Take Before Trying the Keto Diet | Everyday ...

<https://www.everydayhealth.com/diet.../ketogenic-diet/steps-beginners-should-take-be...> ▼

Jan 23, 2018 - Before trying the **ketogenic diet**, you'll need to take a few steps, including knowing what to eat and avoid, embracing cooking, and being aware ...

# Example on Query-focused Summarization

- 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 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 (a paragraph)
  - outline (bullet points)
  - headline (one sentence)

- **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

# Example: Scientific article summarization

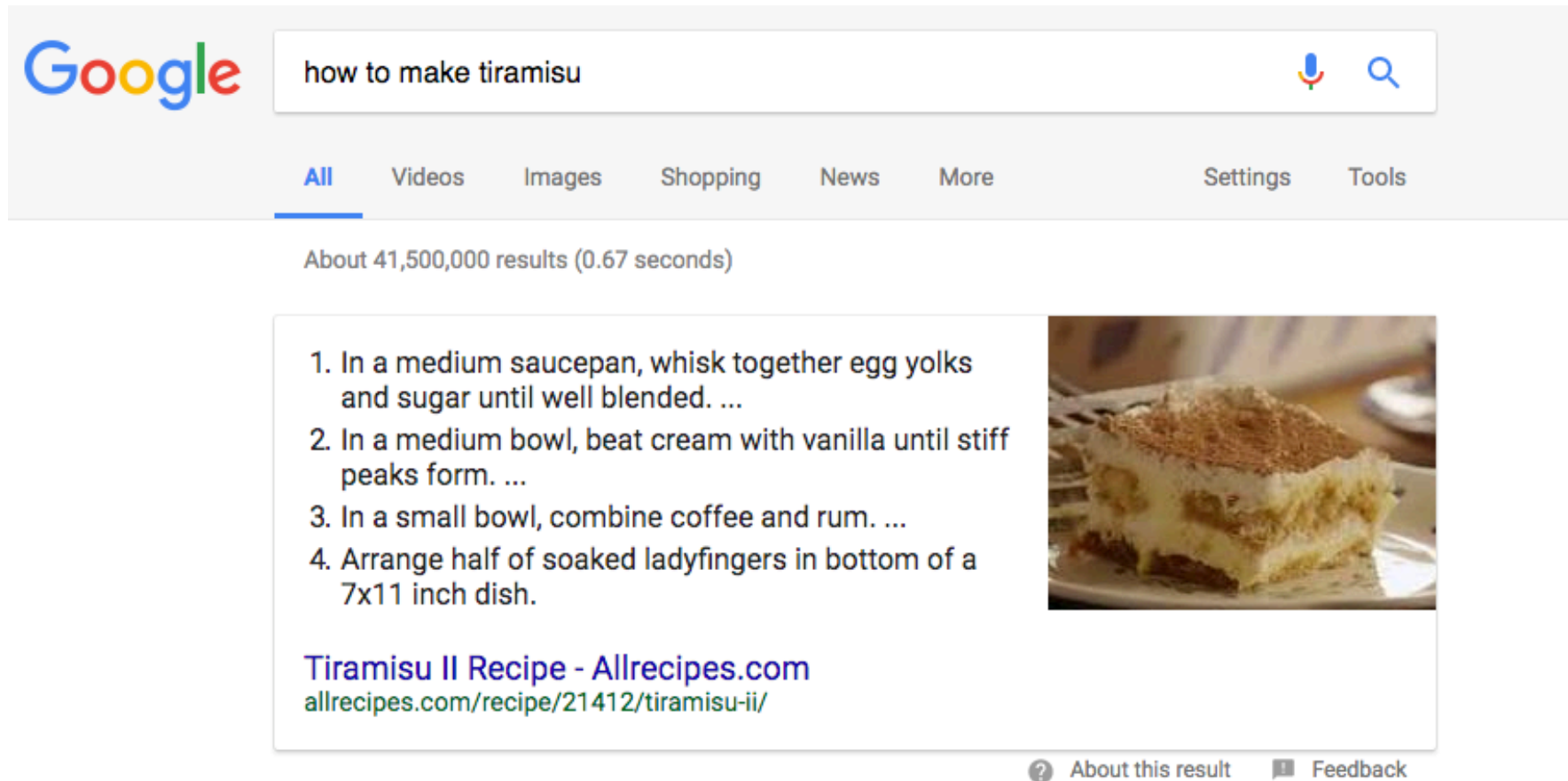
- Single-document summarization task:
  - Not only what the article is about, but also how it relates to work it cites → *summarize the article with regard to prior work*
- Multi-document summarization task:
  - Determine which approaches are criticized and which are supported → *summarization articles that cite a given article*
    - 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 or Search Engine: Featured Snippets

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



The image shows a Google search interface. The search bar contains the text "how to make tiramisu". Below the search bar, there are navigation tabs for "All", "Videos", "Images", "Shopping", "News", and "More", along with "Settings" and "Tools". The search results show "About 41,500,000 results (0.67 seconds)". A featured snippet is displayed, containing a list of four steps for making tiramisu. To the right of the list is a photograph of a slice of tiramisu. Below the list, the source is cited as "Tiramisu II Recipe - Allrecipes.com" with the URL "allrecipes.com/recipe/21412/tiramisu-ii/". At the bottom right, there are links for "About this result" and "Feedback".

Google

how to make tiramisu

All Videos Images Shopping News More Settings Tools

About 41,500,000 results (0.67 seconds)

1. In a medium saucepan, whisk together egg yolks and sugar until well blended. ...
2. In a medium bowl, beat cream with vanilla until stiff peaks form. ...
3. In a small bowl, combine coffee and rum. ...
4. Arrange half of soaked ladyfingers in bottom of a 7x11 inch dish.

**Tiramisu II Recipe - Allrecipes.com**  
[allrecipes.com/recipe/21412/tiramisu-ii/](https://allrecipes.com/recipe/21412/tiramisu-ii/)

About this result Feedback

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)

[https://en.wikipedia.org/wiki/Automatic\\_summarization](https://en.wikipedia.org/wiki/Automatic_summarization)



About this result



Feedback

# 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: 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

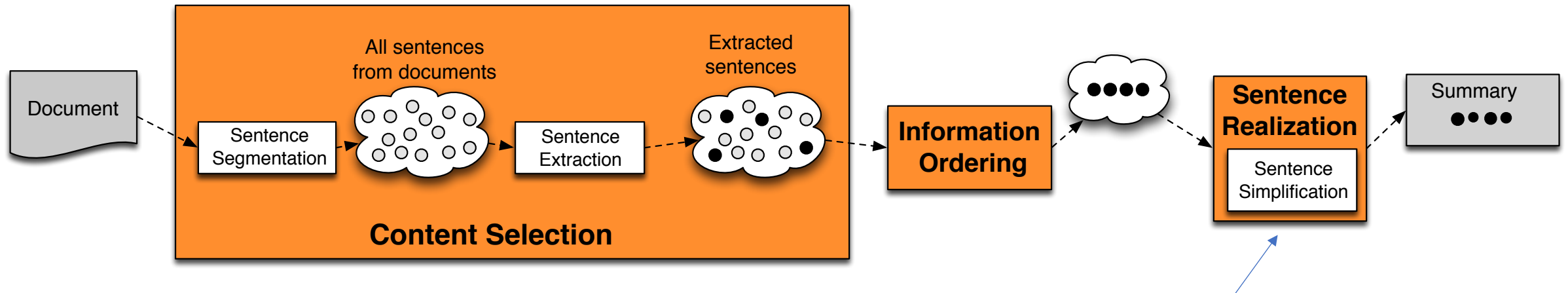
- Input: Congratulations to Australia for seeing sense and dropping the ridiculous policy of not selecting their best players if they are playing overseas.
- Summary: Australia have seen sense by revamping their overseas selection policy.

# Most current systems

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

# Summarization: Three Stages

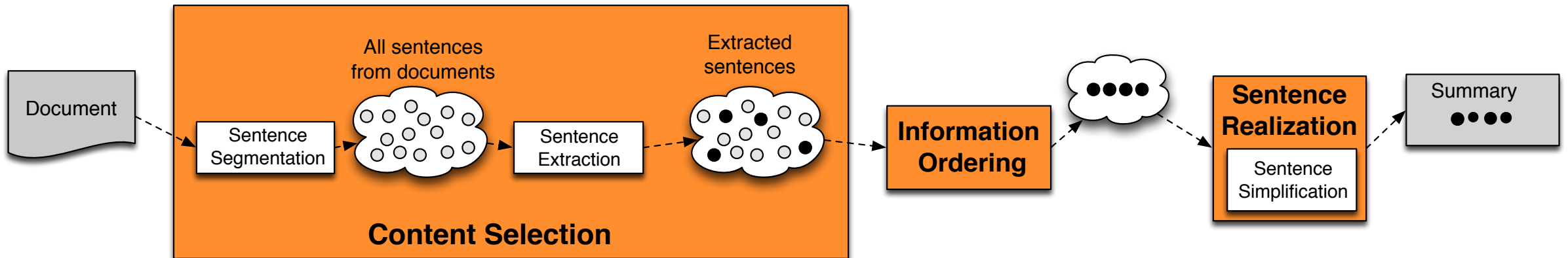
1. **content selection**: choose sentences to extract from the document
2. **information ordering**: choose an order to place them in the summary
3. **sentence realization**: clean up the sentences



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

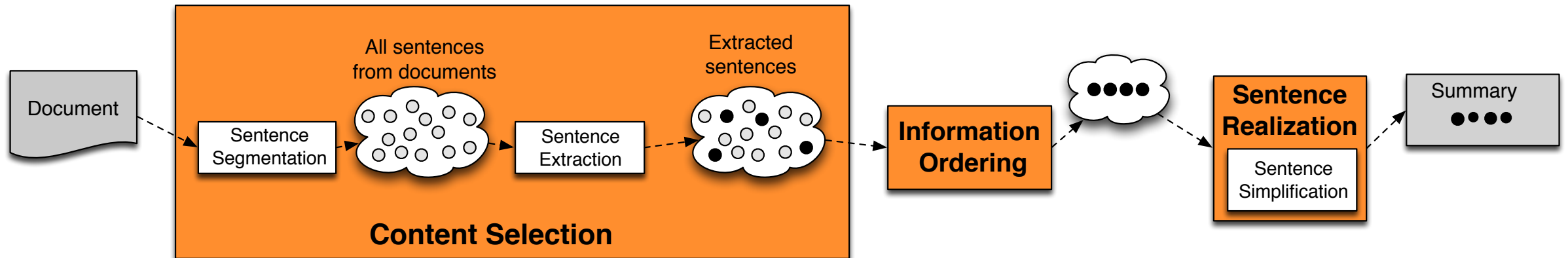
# Basic Summarization Algorithm (extractive)

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

# 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)

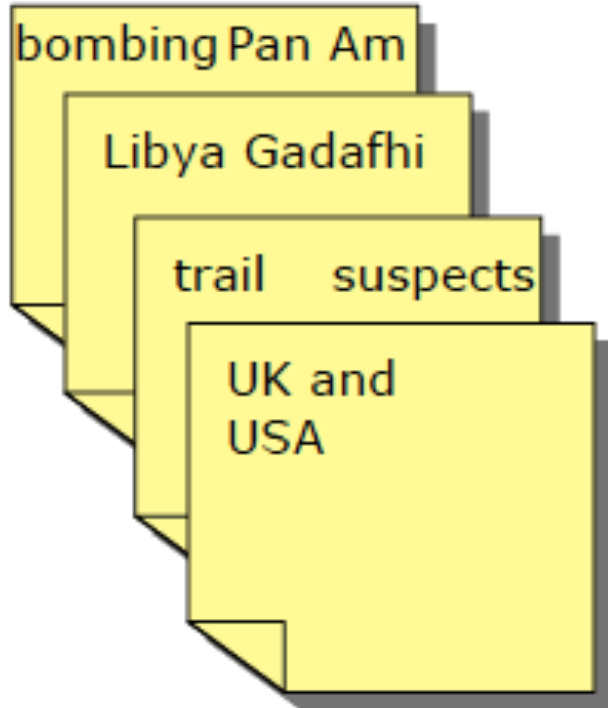
# Frequency as document topic proxy

- 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
  - Proposed by Luhn [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



## 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

HOW?

# 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
- 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

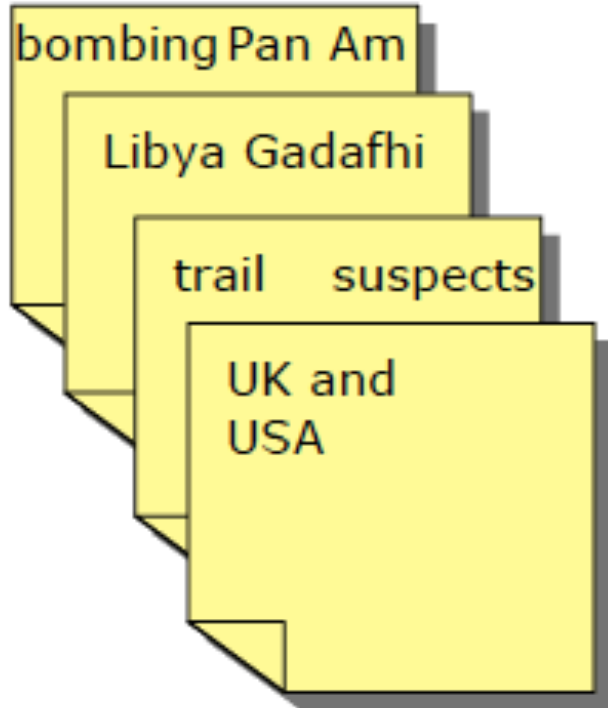
$$Score(S) = \frac{1}{|S|} \sum_{w \in S} p(w)$$

- Update word probabilities for the selected sentence to reduce redundancy

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

- 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

HOW?

# 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!
  - Semantic in general!

# Salient words

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

# 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 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.

# Unsupervised content selection

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

- Topic signatures are assigned with weight of 1

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

# Topic signature-based content selection with queries

Conroy, Schlesinger, and O'Leary 2006

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

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

- 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 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**

# Think: 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.

# 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 (“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?

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



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

# Query-focused Summarization

- Or complex (narrative) 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 choy* (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

## 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
- Stop when desired length

# 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
- Requires more sophisticated NLP techniques

# 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.

# Future Directions: Knowledge-based and Advanced Systems

- Discourse information -> coherent summaries
- Use external lexical resources -> redundancy detection
  - Wordnet, adjective polarity lists, opinion
- Using machine learning models -> neural network and reinforcement learning
- Towards abstractive summarization