Automatic Text Summarization

- **Text summarization** is the process of distilling the most important information from a text to produce a shortened version of a text.

- A **good summary** contains the important and relevant information for a particular task and user.
Summarization applications

• Outlines of any documents
• Abstracts of a scientific article
• Headlines of a news article
• Snippets summarizing a web page on a search engine results page
• Action items or other summaries of a (spoken) business meeting
• Summaries of email threads
• Compressed sentences for producing simplified or compressed text
• Answers to complex questions, constructed by summarizing multiple documents
Automatic summarization

• Single vs multi-document summarization
• Extractive vs abstractive summarization
• Generic vs query-based summarization
Single document summarization

• Given a single document produce a summary
• **Automatic Text Summarizer**

![Automatic Text Summarizer](image-url)
Multi-document summarization

• We are given a group of documents:
  • An email thread
  • A series of news articles on the same topic or event
  • A set of web pages about the same topic or question

• Produce a condensation of the content of the entire group
  • We do not want to summarize each document separately
  • We want to create a cohesive summary that combines the information from each document
Generic summary

• A **generic summary** does not consider a particular user or a particular information need

• The summary simply gives the important information in the document(s)
Query-based summarization

- Also called **query-focused summarization** or **focused summarization** or **topic-based summarization** or **user-focused summarization**
- The summary is generated in a response to a user’s query
- It can be thought of as a kind of longer non-factoid answer to user question
Extractive summarization

• Extractive summarization is formed by selecting (extracting) phrases or sentences from the document
• The summary is generating by pasting together these extracts
Abstractive summarization

- **Abstractive summary** uses (at least in part) different words to describe the contents of the document

- **Abstractive summarization** is a very challenging research task.
Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that this nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we cannot dedicate...we cannot consecrate...we cannot hallow... this ground. The brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. The world will little note nor long remember what we say here, but it can never forget what they did here. It is for us, the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us...that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion; that we here highly resolve that these dead shall not have died in vain; that this nation, under God, shall have a new birth of freedom; and that government of the people, by the people, for the people, shall not perish from the earth. (The Gettisburg Address. Abraham Lincoln, 1863)
An extractive summary

Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field. But the brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. From these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion - that government of the people, by the people, for the people, shall not perish from the earth.
Fourscore and seven years ago our fathers brought forth on this continent a new
nation, conceived in liberty, and dedicated to the proposition that all men are
created equal. Now we are engaged in a great civil war, testing whether that nation,
or any nation so conceived and so dedicated, can long endure. We are met on a great
battle- field of that war. We have come to dedicate a portion of that field as a final
resting-place for those who here gave their lives that this nation might live. It is
altogether fitting and proper that we should do this. But, in a larger sense, we cannot
dedicate...we cannot consecrate...we cannot hallow... this ground. The brave men,
living and dead, who struggled here, have consecrated it far above our poor power
to add or detract. The world will little note nor long remember what we say here, but
it can never forget what they did here. It is for us, the living, rather, to be dedicated
here to the unfinished work which they who fought here have thus far so nobly
advanced. It is rather for us to be here dedicated to the great task remaining before
us...that from these honored dead we take increased devotion to that cause for
which they gave the last full measure of devotion; that we here highly resolve that
these dead shall not have died in vain; that this nation, under God, shall have a new
birth of freedom; and that government of the people, by the people, for the people,
shall not perish from the earth. (The Gettisburg Address. Abraham Lincoln, 1863)
An abstractive summary

This speech by Abraham Lincoln commemorates soldiers who laid down their lives in the Battle of Gettysburg. It reminds the troops that it is the future of freedom in America that they are fighting for.
Basic components of a summarization system

1. **Content selection**: What information to select from the document(s) we are summarizing?

2. **Information ordering**: How to order and structure the extracted units?

3. **Sentence realization**: What kind of clean up to perform on the extracted units so they are fluent in their new context?
Plan

• Single document extractive summarization
• Multi-document extractive summarization
• Query-based extractive summarization
• Neural summarization architectures (both extractive and abstractive)
• Summarization evaluation
Single document summarization
Single document summarization

1. **Content selection**: choose sentences to extract from the document
2. **Information ordering**: choose an order to place these 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

H.P. Luhn, 1958. The Automatic Creation of Literature Abstracts

• Intuition dating back to Luhn (1958)
  • Choose sentences that contain salient or informative words

• Approaches for 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 small set of salient words using log-likelihood ratio (LLR)
• Topic signature is a set of salient or signature terms, whose saliency scores are greater than some threshold

• The LLR for a word $w$ is the ratio between:
  • the probability of observing $w$ both in the input and in the background corpus assuming equal probabilities in both corpora
  • And the probability of observing $w$ in both assuming different probabilities in the input and background corpora

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

• Compute the sentence score by summing the weights of its words:

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

• Then order the sentences by the score
• Finally choose N top ranked sentences into the summary
Centroid-based summarization

• LLR thresholded algorithm belongs to a class of centroid-based summarization algorithms

• The set of signature terms can be viewed as a pseudo-sentence that is the centroid of the document

• Other centrality-based content selection methods rely on tf-idf representation
Centrality method based on tf-idf

1. Compute the tf-idf representation for each sentence
2. Compute the cosine similarity between each pair of sentences in the document
3. Assign a centrality score for each sentence:
   \[
   \text{centrality}(x) = \frac{1}{K} \sum_y \text{tf-idf-cosine}(x, y)
   \]
4. Sentences are ranked according to centrality
Supervised content selection

• Given:
  • A labeled training set of summaries for each document

• Align:
  • The sentences in the document with the sentences in the summary

• Extract features

• Train
  • A binary classifier to decide whether to put the sentence in summary or not
Alignment

• **Human summary**: This paper identifies the desirable features of an ideal multisensor gas monitor and lists the different models currently available.

• **Original document sentence**: The present part lists the desirable features and the different models of portable, multisensor gas monitors currently available.

• Align by finding longest common subsequences of non-stop words.

• Align by using minimum edit distance.
Features used in supervised content selection

- The **position** of the sentence in the document
- **Cue phrases**, such as “in summary”, “in conclusion” or “this paper”
- **Word informativeness** as measured by topic signatures
- **Sentence length**
- **Features** evaluating the cohesion of the sentence
Problems with supervised content selection

• Hard to get labeled training data
• Alignment is difficult (unless the labelled summaries consist of full sentences extracted from the training text)
• Performance is not better than unsupervised algorithms
• Thus in practice, unsupervised content selection is more common
Sentence simplification

• **Original sentence**: When it arrives sometime new 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.
Sentence simplification

• **Original sentence**: 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.

• **Simplified sentence by humans**: The V-chip will give parents a device to block out programs they don’t want their children to see.
Sentence simplification

Use a set of rules on parsed sentences to remove the following:

• **Appositives**: Rajam, 28, an artist who was living at the time in Philadelphia, found the inspiration in the back of the 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 to a sustainable number.

• **Initial adverbials**: “For example”, “On the other hand”, “As a matter of fact”, “At this point”
Multi-document summarization
Multi-document summarization

1. Content selection
2. Information ordering
3. Sentence realization
Maximal Marginal Relevance (MMR)

- An iterative method for content selection from multiple documents
- Iteratively (greedily) choose the best sentence to insert in the summary so far
- The added sentence must be **novel**: minimally redundant with the summary so far:
  - Low cosine similarity to the summary

\[ \text{MMR penalization factor}(s) = \lambda \max_{s_i \in \text{summary}} \text{sim}(s, s_i) \]

- Repeat until the desired number of sentences have been added to the summary
Clustering based content selection

1. Cluster all sentences in the documents
2. Choose a single sentence (centroid) to the summary
Sentence simplification

• Perform sentence simplification during the content selection phase
• Produce several versions of simplified sentences

Former Democratic National Committee finance director Richard Sullivan faced more pointed questioning from Republicans during his second day on the witness stand in the Senate’s fundraising investigation.

• Richard Sullivan faced pointed questioning
• Richard Sullivan faced pointed questioning from Republicans
• Richard Sullivan faced pointed questioning from Republicans during day on stand in Senate fundraising investigation
• Richard Sullivan faced pointed questioning from Republicans in Senate fundraising investigation
Information ordering

• Chronological ordering
  • Order sentences by the date of the document (for summarizing news)

• Coherence
  • Choose ordering that make neighboring sentences similar (by cosine)
  • Choose orderings in which neighboring sentences discuss the same entity
Original summary

Presidential advisers do not blame O’Neill, but they’ve long recognized that a shakeup of the economic team would help indicate Bush was doing everything he could to improve matters. U.S. President George W. Bush pushed out Treasury Secretary Paul O’Neill and top economic adviser Lawrence Lindsey on Friday, launching the first shake-up of his administration to tackle the ailing economy before the 2004 election campaign.
Sentence realization

• Apply coreference resolution, extract names and perform a simple clean-up with rules:
  • Use the **full name** at the first mention, and just the **last name** at subsequent mentions
  • Use a **modified** form for the first mention, but remove appositives or premodifiers from any subsequent mentions
Presidential advisers do not blame Treasury Secretary Paul O’Neill, but they’ve long recognized that a shakeup of the economic team would help indicate U.S. President George W. Bush was doing everything he could to improve matters. Bush pushed out O’Neill and White House economic adviser Lawrence Lindsey on Friday, launching the first shake-up of his administration to tackle the ailing economy before the 2004 election campaign.
Query-based summarization
Query-based summarization

Google search results contain:

• A snippet summarizing the web page for a query
• Title of the web page
• Address / link
Query-based summarization - baseline

• Just take one or few first sentences
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

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

• **Answer:** 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?
Query-based 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-based multi-document summarization
MMR for query-based summarization

• Iteratively (greedily) choose the best sentence to insert in the summary so far so that the added sentence would be:
  • **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

\[
\hat{s}_{\text{MMR}} = \max_{s \in D} \lambda \text{sim}(s, Q) - (1 - \lambda) \max_{s' \in S} \text{sim}(s, s')
\]
Query-based content selection with LLR

• Choose words that are informative either
  • By log-likelihood ratio (LLR)
  • Or by appearing in the query

\[
weight(w_i) = \begin{cases} 
1, & \text{if } -2 \log \lambda(w_i) > 10 \\
1, & \text{if } w_i \in \text{query} \\
0, & \text{otherwise}
\end{cases}
\]
LLR+MMR

• Choosing informative yet non-redundant sentences

1. Score each sentence based on LLR
2. Include the sentence with highest score in the summary
3. Iteratively add high-scoring sentences into the summary that are not redundant with summary so far
Domain-specific information extraction method

• A good **biography** of a person contains:
  • A person’s **birth/death, fame factor, education, nationality** and so on

• A good **definition** contains:
  • **Genus** or **hypernym, species, synonyms, subtypes**

• A **medical answer** about a drug’s use contains:
  • The **problem** (the medical condition)
  • The **intervention** (the drug or procedure)
  • The **outcome** (the result of the study)
## Domain-specific information extraction method

<table>
<thead>
<tr>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>genus</strong></td>
</tr>
<tr>
<td><strong>species</strong></td>
</tr>
<tr>
<td><strong>synonym</strong></td>
</tr>
<tr>
<td><strong>subtype</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Biography</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>dates</strong></td>
</tr>
<tr>
<td><strong>nationality</strong></td>
</tr>
<tr>
<td><strong>education</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drug efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>population</strong></td>
</tr>
<tr>
<td><strong>problem</strong></td>
</tr>
<tr>
<td><strong>intervention</strong></td>
</tr>
<tr>
<td><strong>outcome</strong></td>
</tr>
</tbody>
</table>
"What is the Hajj?"
(Ndocs=20, Len=8)

The Hajj, or pilgrimage to Makkah [Mecca], is the central duty of Islam. More than two million Muslims are expected to take the Hajj this year. Muslims must perform the hajj at least once in their lifetime if physically and financially able. The Hajj is a milestone event in a Muslim's life. The annual hajj begins in the twelfth month of the Islamic year (which is lunar, not solar, so that hajj and Ramadan fall sometimes in summer, sometimes in winter). The Hajj is a week-long pilgrimage that begins in the 12th month of the Islamic lunar calendar. Another ceremony, which was not connected with the rites of the Ka'ba before the rise of Islam, is the Hajj, the annual pilgrimage to 'Arafat, about two miles east of Mecca, toward Mina…
Neural architectures for summarization
Extractive summarization with NN
Cheng and Lapata, 2016. Neural Summarization by Extracting Sentences and Words
Abstractive summarization

• Generating a headline or a short summary consisting of a few sentences that captures the salient ideas of an article or a passage
• Similar to machine translation in some sense: encode the input --> generate the summary
• **There are differences:**
  • Summary is typically short and does not depend on the length of the input
  • Compress the document in a **lossy** manner vs the lossless compression of MT
  • In MT, there is a strong tendency of the one-to-one word-level alignment, in summarization it is less obvious
Abstractive summarization with NN
Nallapati et al., 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond

• Attentional encoder-decoder with feature-rich encoder
Abstractive summarization with NN
Nallapati et al., 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond

• Switching between generation/pointer for generating output
Abstractive summarization with NN

Nallapati et al., 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond

- Hierarchical encoder with hierarchical attention
Evaluating summaries
ROUGE

• Recall Oriented Understudy for Gisting Evaluation (Lin and Hovy, 2003)

• Intrinsic method for automatically evaluating summaries
  • Based on BLEU (a metric used for machine translation)
  • Not as good as human evaluation
  • But much more convenient
Computing ROUGE

• For each test document \( D \) ask \( N \) humans to produce a set of reference summaries \( R \)
• Run the system to produce the automatic summary \( X \) for the document \( D \)
• What percentage of the bigrams in reference summaries \( R \) appear in \( X \)?

\[
\text{ROUGE-2} = \frac{\sum_{S \in R} \sum_{w_i w_j \in S} \min(\text{count}(w_i w_j, X), \text{count}(w_i w_j, S))}{\sum_{S \in R} \sum_{w_i w_j \in S} \text{count}(w_i w_j, S)}
\]
A ROUGE example

• **Query:** What is water spinach?
  • Human1: Water spinach is a green leafy vegetable grown in the tropics.
  • Human2: Water spinach is a semi-aquatic tropical plant grown as a vegetable.
  • Human3: 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 =
A ROUGE example

• **Query:** What is water spinach?
  
  • Human1: Water spinach is a green leafy vegetable grown in the tropics.
  • Human2: Water spinach is a semi-aquatic tropical plant grown as a vegetable.
  • Human3: 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 = \(\frac{3 + 3 + 6}{10 + 10 + 9} = \frac{12}{29} = 0.41\)
ROUGE variants

• **ROUGE-N** – uses N-gram overlap

• **ROUGE-L** – measures the longest common substring between reference and candidate summaries

• **ROUGE-S** – measures the number of *skip bigrams* between reference and candidate summaries
Pyramid Method

• Human-assisted evaluation method, focusses more on meaning

• Humans label the **Summary Content Units** in each reference and candidate summary and then the overlap is computed

• **Summary Content Units**: sub-sentential semantic units which roughly correspond to propositions or coherent pieces of propositions
Summary Content Units

A1. The industrial espionage case involving GM and VW 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 along with ensuing events, 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 eccentric and visionary Basque-born procurement chief Jose Ignacio Lopez de Arriortura and seven of Lopez’s business colleagues.

E1. On March 16, 1993, with Japanese car import quotas to Europe expiring in two years, renowned cost-cutter, Agnacio Lopez De Arriortura, left his job as head of purchasing at General Motor’s Opel, Germany, to become Volkswagen’s Purchasing and Production director.

F3. In March 1993, Lopez and seven other GM executives moved to VW overnight.
Summary Content Units

**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 GMs ... Jose Ignacio Lopez
E1. Agnacio Lopez De Arriortura, left his job ... at General Motors Opel
... to become Volkswagens ... 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
Baselines

• Random sentences baseline
• Leading sentences baseline
Conclusion

• Different types of summarization tasks:
  • Single vs multi-document summarization
  • Generic vs query–based summarization
  • Extractive vs abstractive summarization

• Main steps in extractive summarization:
  • Content selection: tf-idf, LLR, MMR
  • Sentence reordering: chronological, based on similarity
  • Sentence realization: rules for sentence simplification

• Evaluation methods:
  • ROUGE (automatic), Pyramid (human evaluation)