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
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
- Classroom
- Talk Shows
- Chat
- Broadcast News
- Radio News
• “Summaries as short as 17% of the full text length speed up decision making twice, with no significant degradation in accuracy.”

• “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 systems - is the system working?
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
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: Snippets

• Create **snippets** summarizing a web page for a query
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
Neural Abstractive Text 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.
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
Snippets: query-focused summaries

Was cast-metal movable type invented in Korea?

About 591,000 results (0.14 seconds)

Movable type - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Movable_type
Jump to Metal movable type: Transition from wood type to metal type occurred in 1234 ... The following description of the Korean font casting ... In the early fifteenth century, however, the Koreans invented a form of movable type that has ...

History of printing in East Asia - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/History_of_printing_in_East_Asia
The following description of the Korean font casting process was recorded by the ... While metal movable type printing was invented in Korea and the oldest ...

Korea, 1000–1400 A.D. | Heilbrunn Timeline of Art History | The ...
www.metmuseum.org/toah/hd/?period=07&region=ek
The invention and use of cast-metal movable type in Korea in the early thirteenth century predates by two centuries Gutenberg's invention of metal movable type ...
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]
LIBYA REFUSES TO SURRENDER TWO PAN AM BOMBING SUSPECTS

WORD PROBABILITY TABLE

<table>
<thead>
<tr>
<th>Word</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>pan</td>
<td>0.0798</td>
</tr>
<tr>
<td>am</td>
<td>0.0825</td>
</tr>
<tr>
<td>libya</td>
<td>0.0096</td>
</tr>
<tr>
<td>suspects</td>
<td>0.0341</td>
</tr>
<tr>
<td>gadafhi</td>
<td>0.0911</td>
</tr>
<tr>
<td>trail</td>
<td>0.0002</td>
</tr>
<tr>
<td>....</td>
<td></td>
</tr>
<tr>
<td>usa</td>
<td>0.0007</td>
</tr>
</tbody>
</table>

HOW?
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
  \[
  Scord(S) = \frac{1}{|S|} \sum_{w \in S} p(w)
  \]
- Update word probabilities for the selected sentence to reduce redundancy
  \[
  p^{\text{new}}(w) = p^{\text{old}}(w) \cdot p^{\text{old}}(w)
  \]
- Repeat until desired summary length
Obvious shortcomings of the pure frequency approaches

• Does not take account of 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
Topic words (topic signatures)

- 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

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

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**

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1$: $P(t</td>
<td>T) = P(t</td>
</tr>
<tr>
<td>$H_2$: $P(t</td>
<td>T) = p_1$ and $P(t</td>
</tr>
</tbody>
</table>
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

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

\[-2 \log \lambda \text{ has a known statistical distribution: chi-square}\]

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

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

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

(could learn more complex weights)

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

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

- **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?

\[
\text{ROUGE} - 2 = \frac{\sum_{s \in \{\text{RefSummaries}\}} \sum_{\text{bigrams } i \in S} \min(\text{count}(i, X), \text{count}(i, S))}{\sum_{s \in \{\text{RefSummaries}\}} \sum_{\text{bigrams } i \in S} \text{count}(i, S)}
\]
A ROUGE example:

Q: “What is water spinach?”

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 =

\[
\frac{3 + 3 + 6}{10 + 9 + 9} = \frac{12}{28} = .43
\]
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

- a
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)

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

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

• 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.
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