Algorithms for NLP

Summarization

Chan Young Park – CMU
Slides adapted from: Dan Jurafsky – Stanford
Piji Li – Tencent AI Lab
Goal: produce an abridged version of a text that contains information that is *important* or *relevant* to a user.
Text Summarization

- Summarization Applications
  - outlines or abstracts of any document, article, etc
  - summaries of email threads
  - action items from a meeting
  - simplifying text by compressing sentences
Categories

- **Input**
  - Single-Document Summarization (SDS)
  - Multiple-Document Summarization (MDS)

- **Output**
  - Extractive
  - Abstractive

- **Focus**
  - Generic
  - Query-focused summarization

- **Machine learning methods:**
  - Supervised
  - Unsupervised
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
Cambodian leader Hun Sen on Friday rejected opposition parties' demands for talks outside the country, accusing them of trying to "internationalize" the political crisis. Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election negotiations between the two opposition groups and Hun Sen’s party to form a new government failed.

Opposition leaders Prince Norodom Ranariddh and Sam Rainsy, citing Hun Sen’s threats to arrest opposition figures after two alleged attempts on his life, said they could not negotiate freely in Cambodia and called for talks at Sihanouk’s residence in Beijing. Hun Sen, however, rejected that.

I would like to make it clear that all meetings related to Cambodian affairs must be conducted in the Kingdom of Cambodia," Hun Sen told reporters after a Cabinet meeting on Friday. "No-one should internationalize Cambodian affairs. It is detrimental to the sovereignty of Cambodia," he said. Hun Sen's Cambodian People’s Party won 64 of the 122 parliamentary seats in July’s elections, short of the two-thirds majority needed to form a government on its own. Ranariddh and Sam Rainsy have charged that Hun Sen’s victory in the elections was achieved through widespread fraud. They have demanded a thorough investigation into their election complaints as a precondition for their cooperation in getting the national assembly moving and a new government formed.

Figure 1: Single-document summarization.
Multiple-document Summarization

Documents

Fingerprints and photos of two men who boarded the doomed Malaysia Airlines passenger jet are being sent to U.S. authorities so they can be compared against records of known terrorists and criminals. The cause of the plane's disappearance has baffled investigators and they have not said that they believed that terrorism was involved, but they are also not ruling anything out. The investigation into the disappearance of the jetliner with 239 passengers and crew has centered so far around the fact that two passengers used passports stolen in Thailand from an Austrian and an Italian. The plane which left Kuala Lumpur, Malaysia, was headed for Beijing. Three of the passengers, one adult and two children, were American. ...

(CNN) -- A delegation of painters and calligraphers, a group of Buddhists returning from a religious gathering in Kuala Lumpur, a three-generation family, nine senior travelers and five toddlers. Most of the 227 passengers on board missing Malaysia Airlines Flight 370 were Chinese, according to the airline's flight manifest. The 12 missing crew members on the flight that disappeared early Saturday were Malaysian. The airline's list showed the passengers hailed from 14 countries, but later it was learned that two people named on the manifest -- an Austrian and an Italian -- whose passports had been stolen were not aboard the plane. The plane was carrying five children under 5 years old, the airline said. ...

::

Vietnamese aircraft spotted what they suspected was one of the doors belonging to the ill-fated Malaysia Airlines Flight MH370 on Sunday, as troubling questions emerged about how two passengers managed to board the Boeing 777 using stolen passports. The discovery comes as officials consider the possibility that the plane disintegrated mid-flight, a senior source told Reuters. The state-run Thanh Nien newspaper cited Lt. Gen. Vo Van Tuan, deputy chief of staff of Vietnam's army, as saying searchers in a low-flying plane had spotted an object suspected of being a door from the missing jet. It was found in waters about 36 miles south of Tho Chu island, in the same area where oil slicks were spotted Saturday. ...

Summary

Flight MH370, carrying 239 people vanished over the South China Sea in less than an hour after taking off from Kuala Lumpur, with two passengers boarded the Boeing 777 using stolen passports. Possible reasons could be an abrupt breakup of the plane or an act of terrorism. The government was determining the "true identities" of the passengers who used the stolen passports. Investigators were trying to determine the path of the plane by analysing civilian and military radar data while ships and aircraft from seven countries scouring the seas around Malaysia and south of Vietnam.

Figure 2: Multi-document summarization for the topic “Malaysia Airlines Disappearance”.
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
Create **snippets** summarizing a web page for a query

Google: 156 characters (about 26 words) plus title and link
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
History of Summarization

- **Since 1950s:**
  - Concept Weight (Luhn, 1958), Centroid (Radev et al., 2004), LexRank (Erkan and Radev, 2004), TextRank (Mihalcea and Tarau, 2004), Sparse Coding (He et al., 2012; Li et al., 2015)
  - Feature+Regression (Min et al., 2012; Wang et al., 2013)
- Most of the summarization methods are extractive.
- Abstractive summarization is full of challenges.
  - Some indirect methods employ sentence fusing (Barzilay and McKeown, 2005) or phrase merging (Bing et al., 2015).
- The indirect strategies will do harm to the linguistic quality of the constructed sentences.
Methods

How to detect salient words/sentences?
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=eak
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


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

- by log-likelihood ratio (LLR)
- or by appearing in the query

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

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

\[
weight(s) = \frac{1}{|S|} \sum_{w \in S} weight(w)
\]
Graph-based Ranking Algorithms

Rada Mihalcea, ACL 2004

- unsupervised sentence extraction

\[
\text{Similarity}(S_i, S_j) = \frac{|W_k | W_k \in S_i \& W_k \in S_j |}{\log(|S_i|)+\log(|S_j|)}
\]

3: BC–HurricaneGilbert, 09–11 339
4: BC–Hurricane Gilbert, 0348
5: Hurricaine Gilbert heads toward Dominican Coast
6: By Ruddy Gonzalez
7: Associated Press Writer
8: Santo Domingo, Dominican Republic (AP)
9: Hurricaine Gilbert Swept towed the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains, and high seas.
10: The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph.
11: "There is no need for alarm," Civil Defense Director Eugenio Cabral said in a television alert shortly after midnight Saturday.
12: Cabral said residents of the province of Barahona should closely follow Gilbert’s movement.
13: An estimated 100,000 people live in the province, including 70,000 in the city of Barahona, about 125 miles west of Santo Domingo.
14: Tropical storm Gilbert formed in the eastern Carribean and strengthened into a hurricane Saturday night.
15: The National Hurricane Center in Miami reported its position at 2 a.m. Sunday at latitude 16.1 north, longitude 67.5 west, about 140 miles south of Ponce, Puerto Rico, and 200 miles southeast of Santo Domingo.
16: The National Weather Service in San Juan, Puerto Rico, said Gilbert was moving westward at 15 mph with a "broad area of cloudiness and heavy weather" rotating around the center of the storm.
17: The weather service issued a flash flood watch for Puerto Rico and the Virgin Islands until at least 6 p.m. Sunday.
18: Strong winds associated with the Gilbert brought coastal flooding, strong southeast winds, and up to 12 feet to Puerto Rico’s south coast.
19: There were no reports on casualties.
20: San Juan, on the north coast, had heavy rains and gusts Saturday, but they subsided during the night.
21: On Saturday, Hurricane Florence was downgraded to a tropical storm, and its remnants pushed inland from the U.S. Gulf Coast.
22: Residents returned home, happy to find little damage from 90 mph winds and sheets of rain.
23: Florence, the sixth named storm of the 1988 Atlantic storm season, was the second hurricane.
24: The first, Debby, reached minimal hurricane strength briefly before hitting the Mexican coast last month.

Figure 1: Sample graph build for sentence extraction from a newspaper article.
Washington (CNN) — President Donald Trump struck a unifying tone Monday as he addressed the devastation in Texas wrought by Hurricane Harvey at the top of a joint news conference with Finland’s president.

"We see neighbor helping neighbor, friend helping friend and stranger helping stranger," Trump said. "We are one American family. We hurt together, we struggle together and believe me, we endure together."

Trump extended his "thoughts and prayers" to those affected by the hurricane and catastrophic flooding that ensued in Texas, and also promised Louisiana residents that the federal government is prepared to help as the tropical storm makes its way toward that state.

"To the people of Texas and Louisiana, we are 100% with you," Trump said from the East Room of the White House.
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
  - alignment difficult
  - performance not better than unsupervised algorithms

- **So in practice:**
  - Unsupervised content selection is more common
Evaluating Summaries: ROUGE
Intrinsic metric for automatically evaluating summaries

- Based on BLEU (a metric used for machine translation)
- 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?

\[
\text{ROUGE-2} = \frac{\sum_{i \in S} \sum_{s \in \text{RefSummaries}} \min(\text{count}(i, X), \text{count}(i, S))}{\sum_{s \in \text{RefSummaries}} \sum_{i \in S} \text{count}(i, S)}
\]
A ROUGE example:
Q: “What is water spinach?”

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

- Human Summaries (Gold)
  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.

- ROUGE-2 = \[ \frac{3 + 3 + 6}{10 + 9 + 9} = \frac{12}{28} = .43 \]
Neural Text Summarization
A neural attention model for abstractive sentence summarization

- Inspired by attention-based seq2seq models (Bahdanau, 2014)

Figure 5: (a) NNLN decoder with additional encoder element. (b) Attention based encoder.
A neural attention model for abstractive sentence summarization  

Rush et al., EMNLP 2015

- Inspired by attention-based seq2seq models (Bahdanau, 2014)

Input \((x_1, \ldots, x_{18})\). First sentence of article:  
russian defense minister ivanov called sunday for the creation of a joint front for combating global terrorism  

Output \((y_1, \ldots, y_8)\). Generated headline:  
russia calls for joint front against terrorism \(\iff g(\text{terrorism}, x, \text{for}, \text{joint}, \text{front}, \text{against})\)

Figure 2: Example input sentence and the generated summary. The score of generating \(y_{i+1}\) (terrorism) is based on the context \(y_c\) (for ... against) as well as the input \(x_1 \ldots x_{18}\). Note that the summary generated is abstractive which makes it possible to generalize (russian defense minister to russia) and paraphrase (for combating to against), in addition to compressing (dropping the creation of), see Jing (2002) for a survey of these editing operations.
Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond

- Implements many tricks (nmt, copy, coverage, hierarchical, external knowledge)

Figure 1: Feature-rich-encoder: We use one embedding
Figure 2: Switching generator/pointer model: When the
Figure 3: Hierarchical encoder with hierarchical attention:
- Implements many tricks (nmt, copy, coverage, hierarchical, external knowledge)

<table>
<thead>
<tr>
<th>#</th>
<th>Model name</th>
<th>Rouge-1</th>
<th>Rouge-2</th>
<th>Rouge-L</th>
<th>Src. copy rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full length F1 on our internal test set</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>words-lvt2k-1sent</td>
<td>34.97</td>
<td>17.17</td>
<td>32.70</td>
<td>75.85</td>
</tr>
<tr>
<td>2</td>
<td>words-lvt2k-2sent</td>
<td>35.73</td>
<td>17.38</td>
<td>33.25</td>
<td>79.54</td>
</tr>
<tr>
<td>3</td>
<td>words-lvt2k-2sent-hieratt</td>
<td>36.05</td>
<td>18.17</td>
<td>33.52</td>
<td>78.52</td>
</tr>
<tr>
<td>4</td>
<td>feats-lvt2k-2sent</td>
<td>35.90</td>
<td>17.57</td>
<td>33.38</td>
<td>78.92</td>
</tr>
<tr>
<td>5</td>
<td>feats-lvt2k-2sent-ptr</td>
<td>*36.40</td>
<td>*17.77</td>
<td>*33.71</td>
<td>78.70</td>
</tr>
<tr>
<td></td>
<td>Full length F1 on the test set used by (Rush et al., 2015)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ABS+ (Rush et al., 2015)</td>
<td>29.78</td>
<td>11.89</td>
<td>26.97</td>
<td>91.50</td>
</tr>
<tr>
<td>7</td>
<td>words-lvt2k-1sent</td>
<td>32.67</td>
<td>15.59</td>
<td>30.64</td>
<td>74.57</td>
</tr>
<tr>
<td>8</td>
<td>RAS-Elman (Chopra et al., 2016)</td>
<td>33.78</td>
<td>15.97</td>
<td>31.15</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>words-lvt5k-1sent</td>
<td>*35.30</td>
<td>*16.64</td>
<td>*32.62</td>
<td></td>
</tr>
</tbody>
</table>
Copy Mechanism

- OOV, Extraction
- "Pointer networks" (Vinyals et al., 2015 NIPS)
- "Pointing the Unknown Words” (Gulcehre et al., ACL 2016)
- "Incorporating Copying Mechanism in Sequence-to-Sequence Learning " (Gu et al., ACL 2016)
- "Get To The Point: Summarization with Pointer-Generator Networks " (See et al., ACL 2017)
Pointer Generator Networks

Copy words from the source text

See et al., ACL 2017
Pointer Generator Networks

\[
P_{\text{vocab}} = \text{softmax}(V'(V [s_t, h_t^*] + b) + b')
\]

\[
p_{\text{gen}} = \sigma(w_h^T h_t^* + w_s^T s_t + w_x^T x_t + b_{\text{ptr}})
\]

\[
P(w) = p_{\text{gen}}P_{\text{vocab}}(w) + (1 - p_{\text{gen}})\sum_{i: w_i = w} a_i^t
\]

See et al., ACL 2017
**Article:** andy murray (...) is into the semi-finals of the miami open, but not before getting a scare from 21 year-old austrian dominic thiem, who pushed him to 4-4 in the second set before going down 3-6 6-4, 6-1 in an hour and three quarters. (...)

**Summary:** andy murray defeated dominic thiem 3-6 6-4, 6-1 in an hour and three quarters.

---

**Article:** (...) wayne rooney smashes home during manchester united ’s 3-1 win over aston villa on saturday. (...)

**Summary:** manchester united beat aston villa 3-1 at old trafford on saturday.
Neural Extractive Models

- "SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents." (Nallapati et al., AAAI 2017)
Hybrid approach

- "Bottom-Up Abstractive Summarization" (Gehrmann et al., EMNLP 2018)

Figure 2: Overview of the selection and generation processes described throughout Section 4.
Hybrid approach

- "Bottom-Up Abstractive Summarization" (Gehrmann et al., EMNLP 2018)

Bottom-Up Attention Models for Extractive Abstractive Summarization

- Standard → Train (a) → Abstractive → Predict (a) → Abstractive
- Tagging → Process (b) → Extract (e) → Bottom-Up

Legend:
- DATA
- allenNLP
- OpenNMT
- Results

© Sebastian Gehrmann, 2018
Hybrid approach

- "Bottom-Up Abstractive Summarization" (Gehrmann et al., EMNLP 2018)

<table>
<thead>
<tr>
<th>Method</th>
<th>R-1</th>
<th>R-2</th>
<th>R-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML + RL (Paulus et al., 2017)</td>
<td>39.87</td>
<td>15.82</td>
<td>36.90</td>
</tr>
<tr>
<td>Saliency + Entailment reward (Pasunuru and Bansal, 2018)</td>
<td>40.43</td>
<td>18.00</td>
<td>37.10</td>
</tr>
<tr>
<td>Key information guide network (Li et al., 2018a)</td>
<td>38.95</td>
<td>17.12</td>
<td>35.68</td>
</tr>
<tr>
<td>Inconsistency loss (Hsu et al., 2018)</td>
<td>40.68</td>
<td>17.97</td>
<td>37.13</td>
</tr>
<tr>
<td>Sentence Rewriting (Chen and Bansal, 2018)</td>
<td>40.88</td>
<td>17.80</td>
<td>38.54</td>
</tr>
<tr>
<td>Pointer-Generator (our implementation)</td>
<td>36.25</td>
<td>16.17</td>
<td>33.41</td>
</tr>
<tr>
<td>Pointer-Generator + Coverage Penalty</td>
<td>39.12</td>
<td>17.35</td>
<td>36.12</td>
</tr>
<tr>
<td>CopyTransformer + Coverage Penalty</td>
<td>39.25</td>
<td>17.54</td>
<td>36.45</td>
</tr>
<tr>
<td>Pointer-Generator + Mask Only</td>
<td>37.70</td>
<td>15.63</td>
<td>35.49</td>
</tr>
<tr>
<td>Pointer-Generator + Multi-Task</td>
<td>37.67</td>
<td>15.59</td>
<td>35.47</td>
</tr>
<tr>
<td>Pointer-Generator + DiffMask</td>
<td>38.45</td>
<td>16.88</td>
<td>35.81</td>
</tr>
<tr>
<td>Bottom-Up Summarization</td>
<td>41.22</td>
<td>18.68</td>
<td>38.34</td>
</tr>
<tr>
<td>Bottom-Up Summarization (CopyTransformer)</td>
<td>40.96</td>
<td>18.38</td>
<td>38.16</td>
</tr>
</tbody>
</table>
Other lines of research

- **Coverage Mechanism**
  - “Modeling Coverage for Neural Machine Translation” (Tu et al., 2016 ACL)

- **Graph-based attentional neural model**
  - “Abstractive document summarization with a graph-based attentional neural model” (Tan et al., ACL 2017)

- **Reinforcement Learning**
  - “A deep reinforced model for abstractive summarization.” (Paulus et al., ICLR 2018)
Conclusion
Conclusion

- Salient Detection
  - How to detect important/relevant words or sentences?

- Remaining Challenges
  - Long text abstractive summarization
  - Abstractive multi-document summarization