NLP
Introduction to NLP

Summarization Techniques 1/3
Baxendale (1958)

• Positional method
  – Analysis of 200 paragraphs
  – Pick the first and last sentences of the paragraph
  – That’s where the topic sentences are located
  – Naïve but decent approach
Luhn (1958)

- Technical documents
- Stemming
- Stop words are removed
- Frequency of content terms
Luhn (1958)

- Sentence-level significance factor
- Look for concentrations of salient content terms

Score = $4^2/7 \approx 2.3$
Edmundson (1969)

- Technical documents
- Position and frequency
- Cue words (bonus and stigma words)
  - Significant, hardly, impossible
- Document structure
  - Is the sentence a title or heading or right under one of these
- Linear combination of the four features
  \[ \alpha_1 C + \alpha_2 K + \alpha_3 T + \alpha_4 L \]
Frump (deJong 1979, 1982)

- Knowledge-based
- Slot-filling based on UPI news stories
- Based on 50 sketchy scripts
- Inputs matched to scripts based on manually selected keywords
- Difficult to port to other domains
- Missing scripts for many inputs
Frump

$\text{demonstration script}$

• The demonstrators arrive at the demonstration location.
• The demonstrators march.
• Police arrive on the scene.
• The demonstrators communicate with the target of the demonstration.
• The demonstrators attack the target of the demonstration.
• The demonstrators attack the police.

G. DeJong (1979) FRUMP: Fast Reading Understanding and Memory Program
Paice (1990)

- Survey up to 1990
- Techniques that (mostly) failed
  - Syntactic criteria (Earl 1970)
  - Indicator phrases
- Problems with extracts
  - Lack of balance
  - Lack of cohesion
Paice (1990)

• Lack of balance
  – later approaches based on text rhetorical structure

• Lack of cohesion
  – anaphoric reference
  – lexical or definite reference
  – rhetorical connectives
  – recognition of anaphors [Liddy et al. 87]
  – Example: “that” is
    • *nonanaphoric* if preceded by a research-verb (e.g., “demonstrat-”),
    • *nonanaphoric* if followed by a pronoun, article, quantifier,…,
    • *external* if no later than 10th word,
      else
    • *internal*
• ANES: commercial news from 41 publications
• “Lead” achieves acceptability of 90% vs. 74.4% for “intelligent” summaries
• 20,997 documents
• words selected based on tf*idf (term frequency * inverse document frequency)
• sentence-based features:
  – signature words
  – location
  – anaphora words
  – length of abstract
• Sentences with no signature words are included if between two selected sentences
• Evaluation done at 60, 150, and 250 word length
• Non-task-driven evaluation:
  – “Most summaries judged less-than-perfect would not be detectable as such to a user”
Kupiec et al. (1995)

• First trainable method
  – 20% extract
  – 188 documents from scientific journals
  – Naïve Bayes classifier

• New features
  – Sentence length ($|S|>5$)
  – Presence of uppercase words (except common acronyms)
  – Thematic words
  – Set of 26 manually fixed phrases
  – Sentence position in paragraph
Kupiec et al. (1995)

- Uses Naïve Bayes classifier

\[
P(s \in S \mid F_1, F_2, \ldots F_k) = \frac{P(F_1, F_2, \ldots F_k \mid s \in S)P(s \in S)}{P(F_1, F_2, \ldots F_k)}
\]

- Assuming statistical independence

\[
P(s \in S \mid F_1, F_2, \ldots F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid s \in S)P(s \in S)}{\prod_{j=1}^{k} P(F_j)}
\]
Kupiec et al. (1995)

• **Performance:**
  – For 25% summaries, 84% precision
  – For smaller summaries, 74% improvement over Lead
Summons (McKeown & Radev 1995)

• First work on multi-document summarization
• Approach
  – Knowledge-based
  – Information extraction (MUC templates)
  – Text generation
NICOSIA, Cyprus (AP) – Two bombs exploded near government ministries in Baghdad, but there was no immediate word of any casualties, Iraqi dissidents reported Friday. There was no independent confirmation of the claims by the Iraqi National Congress. Iraq’s state-controlled media have not mentioned any bombings.

Multiple sources and disagreement

Explicit mentioning of “no information”. 

Summons (McKeown and Radev 1995)
### Summons

<table>
<thead>
<tr>
<th>MESSAGE: ID</th>
<th>TST3-MUC4-0010</th>
</tr>
</thead>
<tbody>
<tr>
<td>MESSAGE: TEMPLATE</td>
<td>2</td>
</tr>
<tr>
<td>INCIDENT: DATE</td>
<td>30 OCT 89</td>
</tr>
<tr>
<td>INCIDENT: LOCATION</td>
<td>EL SALVADOR</td>
</tr>
<tr>
<td>INCIDENT: TYPE</td>
<td>ATTACK</td>
</tr>
<tr>
<td>INCIDENT: STAGE OF EXECUTION</td>
<td>ACCOMPLISHED</td>
</tr>
<tr>
<td>INCIDENT: INSTRUMENT ID</td>
<td></td>
</tr>
<tr>
<td>INCIDENT: INSTRUMENT TYPE</td>
<td></td>
</tr>
<tr>
<td>PERP: INCIDENT CATEGORY</td>
<td>TERRORIST ACT</td>
</tr>
<tr>
<td>PERP: INDIVIDUAL ID</td>
<td>&quot;TERRORIST&quot;</td>
</tr>
<tr>
<td>PERP: ORGANIZATION ID</td>
<td>&quot;THE FMLN&quot;</td>
</tr>
<tr>
<td>PERP: ORG. CONFIDENCE</td>
<td>REPORTED: &quot;THE FMLN&quot;</td>
</tr>
<tr>
<td>PHYS TGT: ID</td>
<td></td>
</tr>
<tr>
<td>PHYS TGT: TYPE</td>
<td></td>
</tr>
<tr>
<td>PHYS TGT: NUMBER</td>
<td></td>
</tr>
<tr>
<td>PHYS TGT: FOREIGN NATION</td>
<td></td>
</tr>
<tr>
<td>PHYS TGT: EFFECT OF INCIDENT</td>
<td></td>
</tr>
<tr>
<td>PHYS TGT: TOTAL NUMBER</td>
<td></td>
</tr>
<tr>
<td>HUM TGT: NAME</td>
<td></td>
</tr>
<tr>
<td>HUM TGT: DESCRIPTION</td>
<td>&quot;1 CIVILIAN&quot;</td>
</tr>
<tr>
<td>HUM TGT: TYPE</td>
<td>CIVILIAN: &quot;1 CIVILIAN&quot;</td>
</tr>
<tr>
<td>HUM TGT: NUMBER</td>
<td>1: &quot;1 CIVILIAN&quot;</td>
</tr>
<tr>
<td>HUM TGT: FOREIGN NATION</td>
<td></td>
</tr>
<tr>
<td>HUM TGT: EFFECT OF INCIDENT</td>
<td>DEATH: &quot;1 CIVILIAN&quot;</td>
</tr>
<tr>
<td>HUM TGT: TOTAL NUMBER</td>
<td></td>
</tr>
</tbody>
</table>
Input: Cluster of templates

\[ T_1 \quad T_2 \quad \ldots \quad T_m \]

Conceptual combiner

- Domain ontology

Combiner

- Planning operators

Paragraph planner

Linguistic realizer

- Lexicon

Sentence planner

Lexical chooser

Sentence generator

OUTPUT: Base summary

Domain ontology
### Summons

<table>
<thead>
<tr>
<th>MESSAGE ID</th>
<th>TST-REU-0001</th>
<th>SECSOURCE: SOURCE</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>SECSOURCE: DATE</td>
<td>March 3, 1996 11:30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRIMSOURCE: SOURCE</td>
<td>March 3, 1996</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCIDENT: DATE</td>
<td>Jerusalem</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCIDENT: LOCATION</td>
<td>Bombing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCIDENT: TYPE</td>
<td>“killed: 18”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUM TGT: NUMBER</td>
<td>“wounded: 10”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PERP: ORGANIZATION ID</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MESSAGE ID</th>
<th>TST-REU-0002</th>
<th>SECSOURCE: SOURCE</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>SECSOURCE: DATE</td>
<td>March 4, 1996 07:20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRIMSOURCE: SOURCE</td>
<td>March 4, 1996</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCIDENT: DATE</td>
<td>Tel Aviv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCIDENT: LOCATION</td>
<td>Bombing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCIDENT: TYPE</td>
<td>“killed: at least 10”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUM TGT: NUMBER</td>
<td>“wounded: more than 100”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PERP: ORGANIZATION ID</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MESSAGE ID</th>
<th>TST-REU-0003</th>
<th>SECSOURCE: SOURCE</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>SECSOURCE: DATE</td>
<td>March 4, 1996 14:20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRIMSOURCE: SOURCE</td>
<td>March 4, 1996</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCIDENT: DATE</td>
<td>Tel Aviv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCIDENT: LOCATION</td>
<td>Bombing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCIDENT: TYPE</td>
<td>“killed: at least 13”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUM TGT: NUMBER</td>
<td>“wounded: more than 100”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PERP: ORGANIZATION ID</td>
<td>“Hamas”</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MESSAGE ID</th>
<th>TST-REU-0004</th>
<th>SECSOURCE: SOURCE</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>SECSOURCE: DATE</td>
<td>March 4, 1996 14:30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRIMSOURCE: SOURCE</td>
<td>March 4, 1996</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCIDENT: DATE</td>
<td>Tel Aviv</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCIDENT: LOCATION</td>
<td>Bombing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCIDENT: TYPE</td>
<td>“killed: at least 12”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUM TGT: NUMBER</td>
<td>“wounded: 105”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PERP: ORGANIZATION ID</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summons

Reuters reported that 18 people were killed on Sunday in a bombing in Jerusalem. The next day, a bomb in Tel Aviv killed at least 10 people and wounded 30 according to Israel radio. Reuters reported that at least 12 people were killed and 105 wounded in the second incident. Later the same day, Reuters reported that Hamas has claimed responsibility for the act.
Summons

• If there are two templates
  AND
  the location is the same
  AND
  the time of the second template is after the time of the first template
  AND
  the source of the first template is different from the source of the second template
  AND
  at least one slot differs
  THEN
  combine the templates using the contradiction operator...
Summons

Change of perspective

Precondition:
The same source reports a change in a small number of slots

March 4th, Reuters reported that a bomb in Tel Aviv killed at least 10 people and wounded 30. Later the same day, Reuters reported that exactly 12 people were actually killed and 105 wounded.
Summons

Contradiction

Precondition:
Different sources report contradictory values for a small number of slots

The afternoon of February 26, 1993, Reuters reported that a suspected bomb killed \textit{at least six people} in the World Trade Center. However, Associated Press announced that \textit{exactly five people} were killed in the blast.
Summons

Refinement
On Monday morning, Reuters announced that a suicide bomber killed at least 10 people in Tel Aviv. In the afternoon, Reuters reported that Hamas claimed responsibility for the act.

Agreement
The morning of March 1st 1994, both UPI and Reuters reported that a man was kidnapped in the Bronx.
According to UPI, three terrorists were arrested in Medellín last Tuesday. Reuters announced that the police arrested two drug traffickers in Bogotá the next day.

A total of five criminals were arrested in Colombia last week.
Mitra/Allen/Salton

- Encyclopedia articles
- Semantic hyperlinks
  - Between pairs of paragraphs with lexical similarity above a threshold
  - Paths linking highly connected paragraphs are more likely to contain information central to the topic of the article
NLP