Text Summarization

Gerald Penn

CSC 401
University of Toronto

http://www.cs.toronto.edu/~gpenncsc401
Text Summarization

Objective: return shortened version of text that includes its main points.

This includes:

- "gisting": just a few words — almost topic classification
- abstracting, e.g., in MS Word
- longer summaries, e.g., 20% of original document size)
- original length (from multiple documents)
Kinds of Summaries

• Text vs. template

• **Perspective**: informative vs. indicative

• **Composition**: extract vs. abstract

• **Orientation**: document vs. query

• **Source**: single vs. multiple document

• **Background**: complete vs. update
Summarization by Extraction

Identify important information, and drop it into summary.

How do we determine importance?

• Position in text, e.g.:
  – first sentence of each paragraph
  – first and last paragraphs of document
  – section headings, captions, etc.
  – varies with genre
  – Hovy-Lin (partial) ordering:
    * WSJ: $T > P1S1 > P1S2 > \cdots$
    * Ziff-Davis: $T > P2S1 > P3S1 > P2S2 > \{P4S1, P5S1, P3S2\} > \cdots$
Summarization by Extraction

Identify important information, and drop it into summary.

How do we determine importance?

- Position in text
- *Indicators*
  - *cues*, e.g.:
    * “in this paper, we show”
    * “in conclusion”
    * “recommend that”
  - *clues (bonus words)*, e.g.:
    * “significantly”
    * “this paper”
  - *stigma words*, e.g.:
    * “hardly”
    * “incidentally”
    * “supported by a grant”
Summarization by Extraction

Identify important information, and drop it into summary.

How do we determine importance?

- Position in text
- **Indicators**
  - *cues*
  - *clues* (bonus words)
  - *stigma words*
  - content words from title
  - **not** tf.idf
Naive Bayes Classification

We can treat summarization as a sequence of binary classification problems: every sentence is either in or out.

Bayes decision rule: choose outcome that is most probable in given context of features:

$$\max\{ P(s \in \text{Summary}|f_1 \ldots f_k), \quad P(s \not\in \text{Summary}|f_1 \ldots f_k) \}$$

$P(o|f_1 \ldots f_k)$ is hard to measure, so we use Bayes’s rule:

$$P(o|f_1 \ldots f_k) = \text{what?}$$
Naive Bayes Classification

We can treat summarization as a \textit{sequence} of \textit{binary} classification problems: every sentence is either \textit{in} or \textit{out}.

Bayes decision rule: choose outcome that is most probable in given context of features:

\[
\max \{ P(s \in \text{Summary}|f_1 \ldots f_k), \quad P(s \not\in \text{Summary}|f_1 \ldots f_k) \} \\
\]

\(P(o|f_1 \ldots f_k)\) is hard to measure, so we use Bayes’s rule:

\[
P(o|f_1 \ldots f_k) = \frac{P(f_1 \ldots f_k|o)P(o)}{P(f_1 \ldots f_k)}
\]

The \textit{Naive Bayes Assumption}: all features of context are conditionally independent. Thus:

\[
P(f_1 \ldots f_k|o) = \prod_{1 \leq j \leq k} P(f_j|o)
\]

And we can use relative frequency in annotated corpora for these:

\[
P(f_j|o) = \frac{C(f_j, o)}{C(o)}
\]
Disadvantages of Summarization by Extraction

- Hard to read, misleading, and/or incoherent, e.g.:
  - lost pronoun antecedents
  - discourse/argument connectives no longer appropriate
- Parts of extracted sentences may be unimportant
  - negation (of clues and stigma words)
  - granularity of sentence-sized extracts
Improvements upon Summarization by Extraction

- Use argument structure to determine importance
- *Cut-and-paste* summarization: use extraction at phrase level to make new sentences
- Summarize multiple documents and use comparisons to boost confidence in importance.
- Task-based evaluation: determine how well summaries work in context. How do people use summaries?