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
Overview summary of Angina

You are at:  Angina
Search:  ○ all documents  ○ within Angina

Get more detailed information on the sections:  [ variant angina: | what is the treatment? | diagnosis | signs and symptoms | what are the symptoms | treatment ]

Synopsis:  Treatment is designed to prevent or reduce ischemia and minimize symptoms. Angina that cannot be controlled by drugs and lifestyle changes may require surgery. Angina attacks usually last for only a few minutes, and most can be relieved by rest. Most often, the discomfort occurs after strenuous physical activity or an emotional upset. A doctor diagnoses angina largely by a person’s description of the symptoms. The underlying cause of angina requires careful medical treatment to prevent a heart attack. Not everyone with ischemia experiences angina. If you experience angina, try to stop the activity that precipitated the attack.

Highlighted differences between the documents:
○ This file (5 minute emergency medicine consult) is close in content to the summary.
○ More information on additional topics which are not included in the summary are available in these files (The American Medical Association family medical guide and The Columbia University College of Physicians and Surgeons complete home medical guide). The topics include "definition" and "what are the risks?"
○ The Merck manual of medical information contains extensive information on the topic.
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?

- *Indicators*, 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
Naïve Bayes and SoftMax

• Broadly, Bayesian probability conceives of probability *not* as frequency of some phenomenon occurring, but rather as an expectation related to our own certainty.

• Given an observation $x$, **Naïve Bayes** simply chooses the class $c \in C$ that maximizes $P(c \mid x)$.
  
  This can be done in many ways.

$$\underset{c}{\text{argmax}} \ P(c \mid x) = \frac{P(c)}{P(x)} P(x \mid c)$$

Estimate the $P(\cdot)$ using Gaussians, or...
Bayesian Classifier

Given features \( \mathbf{x} = [x_1, x_2, \cdots, x_D]^T \)
want to compute class probabilities using Bayes Rule:

\[
\Pr(\text{feature given class}) = \frac{\Pr(\mathbf{x}|c)}{p(c)} \cdot \frac{p(c)}{p(\mathbf{x})}
\]

In words,

Posterior for class = \( \frac{\text{Pr. of feature given class} \times \text{Prior for class}}{\text{Pr. of feature}} \)

To compute \( p(c|\mathbf{x}) \) we need: \( p(\mathbf{x}|c) \) and \( p(c) \).
Independence Assumption

- Naive assumption: The features $x_i$ are conditionally independent given the class $c$.
- Allows us to decompose the joint distribution:
  
  $$p(c, x_1, \ldots, x_D) = p(c) p(x_1|c) \cdots p(x_D|c).$$

- Compact representation of the joint distribution.
  - Prior probability of class:
    $$p(c = 1) = \pi$$
  - Conditional probability of feature given class:
    $$p(x_j = 1|c) = \theta_{jc}$$
Naïve Bayes and SoftMax

Assume $x \in \mathbb{R}^d$, learning a linear decision boundary is tantamount to learning $W \in \mathbb{R}^{C \times d}$.

$$P(\text{Class}|\text{features}) = P(\text{features}|\text{Class}) \cdot P(\text{Class})$$

For all $c \in C$: $f_c = W[c, \ldots] \cdot x = \sum_{i=1}^{d} W[c, i] \cdot x[i]$

Uh oh — $f_c$ can be negative and we want something on $[0,1]$, to be a probability. Solution: Just raise it with an exponent.

Softmax:

$$P(y|x) = \frac{\exp(f_y)}{\sum_{c=1}^{C} \exp(f_c)}$$
Naive Bayes Properties

- An amazingly cheap learning algorithm!
  - **Training time**: Estimate parameters using maximum likelihood.
    - Compute co-occurrence counts of each feature with the labels.
    - Requires only one pass through the data!
  - **Test time**: Apply Bayes’ Rule.
    - Cheap because of the model structure. For more general models, Bayesian inference can be very expensive and/or complicated.
- Analysis easily extends to prob. distributions other than Bernoulli.
- Less accurate in practice compared to discriminative models due to its “naive” independence assumption.
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

$P(o|f_1 \ldots f_k)$ 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 *Naive Bayes Assumption*: all features of context are conditionally independent. Thus:

$$P(f_1 \ldots f_k|o) \equiv \prod_{1 \leq j \leq k} P(f_j|o)$$

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

$$P(f_j|o) = \frac{C(f_j, o)}{C(o)}$$
Disadvantages of Summ. 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 on Summ. by Extraction

- Abstraction
- Use argument structure (???) to determine importance and conservatively synthesize new text
- Summarize multiple documents/background collection and use comparisons to boost confidence in importance.

- Task-based evaluation: determine how well summaries work in context. How do people use summaries?
Remember this?

**NMT: the seq2seq model**

Encoder (RNN) produces an encoding of the source (French) sentence

- The *seq2seq* model is an example of conditioned language model (LM)
- Many variants exist. The classical (vanilla) seq2seq model outlined here
- NMT directly calculates $y^* = \arg \max_y P(y|x)$
- I.e. with our formulation: $E^* = \arg \max_E P(E|F)$

Decoder (RNN) generates target sentence (in English), conditioned on the encoding

Decoder is predicting the next word of the target sentence $y$

Prediction is **conditioned** on the source sentence $x$

$$P(y|x) = P(y_1|x)P(y_2|y_1, x) ... P(y_T|y_1, ... y_{T-1}, x)$$
Attention for Summarization

SEQ2SEQ WITH ATTENTION - GRU

(encoder RNN - GRU)

(towardsdatascience.com)
The trial for one of two men accused *in the beating death of University of Wyoming student Matthew Shepard* will begin with jury selection March 24.

Officials said about 30 people would be questioned.

Henderson and the 21-year-old McKinney both face the death penalty if convicted in a case *that has become a central focus for gay rights activists and others seeking stronger bias crime legislation.*

“There is no guarantee that these laws will stop hate crimes from happening.”

---

**Defendant**

**Charges**

**Why**

**Investigator**

**Sentence**

**Broader Consequences**
Text-to-text Generation

• **Sentence compression** (e.g., Knight and Marcu, 2000)
  
  *Bil Mar Foods Co., a meat processor owned by Sara Lee, announced a recall of certain lots of hot dogs and packaged meat.*

• **Sentence fusion** – merge parts of *similar sentences* (Barzilay and McKeown, 2005; Filippova and Strube, 2008)
  
  *Bil Mar Foods Co. announced a recall of certain lots of hot dogs and packaged meat.*
  
  *The outbreak led to the recall on Tuesday of certain lots of hot dogs and packaged meat produced at the Bil Mar Foods plant.*
Sentence Enhancement

• Traditional fusion: align similar core sentences
  Bil Mar Foods Co. announced a recall of certain lots of hot dogs and packaged meat.
  The outbreak led to the recall on Tuesday of meats produced at the Bil Mar Foods plant.

• Expand with parts of dissimilar in-domain sentences
  This fact has been underscored in the last few months by two unexpected outbreaks of food-borne illness.

• Output:
  The outbreak of food-borne illness led to the recall on Tuesday of certain lots of hot dogs and packaged meat produced at the Bil Mar Foods plant.
Why Sentence Enhancement?

Still a big gap between text-to-text generation methods and what concept-to-text generation aspires to attain

- Locality of context
- Inference
  - Sentence fusion attractive because it doesn’t require deep semantic analysis.
  - *Can we do more with equally little?*
Why Look Beyond Core Sentences

• Automatic summaries *already more central* than human-written gold standard summaries

• **Signature caseframe density** (Cheung and Penn, 2013): a measure of centrality based on method of Lin and Hovy, (2000)
Why Look Beyond Core Sentences

• Automatic summaries *already more central* than human-written gold standard summaries

• **Signature caseframe density** (Cheung and Penn, 2013): a measure of centrality based on method of Lin and Hovy, (2000)
UNSUPERVISED SENTENCE ENHANCEMENT FOR AUTOMATIC SUMMARIZATION

Jackie Chi Kit Cheung and Gerald Penn

University of Toronto
27th October, 2014
Why Look Beyond Core Sentences? A Case Study of the TAC 2010 Corpus

• Where do the semantic predicates in human-written summaries come from?
  • 77% are found in the source text
  • Additional 21% found in in-domain articles
    • 98% of total found!
  • cf., additional 14% found if adding irrelevant articles
  • 51% of source-external sentences use source-internal predicates (often with different arguments)
Why do we look outside the source?

• Source-external predicates are of lower average frequency (95% confidence intervals):

<table>
<thead>
<tr>
<th></th>
<th>Average freq (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source-internal</td>
<td>1.77 (1.57, 2.08)</td>
</tr>
<tr>
<td>Source-external</td>
<td>1.15 (0.99, 1.50)</td>
</tr>
</tbody>
</table>

(Wilcoxon ranked-sums $p < 10^{-17}$)

• Source-external predicates have lower average argument entropy (95% confidence intervals):

<table>
<thead>
<tr>
<th></th>
<th>Arg entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source-internal</td>
<td>7.94 (7.90, 7.97)</td>
</tr>
<tr>
<td>Source-external</td>
<td>7.42 (7.37, 7.48)</td>
</tr>
</tbody>
</table>
Latent-Query Document Summarization

\[
\mathcal{L} = \mathbb{E}_{q_{\phi}(z|x)} \left[ \log p_\theta(y|x, z) \right] + \beta \mathcal{H}(q_{\phi}(z|x)) - \omega \mathcal{H}(r(\hat{z}|x, y), q_{\phi}(z|x))
\]

where \( \mathcal{H}(\cdot) \) denotes posterior entropy and \( \mathcal{H}(\cdot, \cdot) \) denotes cross entropy.

\( (Xu \text{ and Lapata, 2022}) \)