Summarization

Regina Barzilay
EECS Department
MIT
November 9, 2004

Today

- Summarization (content selection, surface realization, evaluation)
- Techniques: alignment, rewriting

Types of Summarization

- Input: speech/text, single-/multi-document, mono-/multilingual
- Output: generic/query-oriented, indicative/informative
- Approach: domain dependent/independent, extraction/generation

Different Ways to Summarize

- Find important information in a text
  - How to define important?
- Learn transformation rules based on training instances
  - We assume that a person selected information in consistent fashion
- Extract certain facts from input, and combine them into a text
  - Appropriate for domain-dependent approaches
Learning to Summarize: Key Issues

- Acquisition of training data
- Content selection
- Content organization and linguistic realization
- Evaluation

Supervised Approaches

- Alignment (trivial for extraction, hard for generation)
- Abstraction
- Classification (standard classifiers — Naive Bayes, SVM, maximum entropy, Boostexter)
- Rewriting (Optional)
Amsterdam is the largest city in the Netherlands and the country’s economic center. It is the official capital of the Netherlands, though The Hague is the home of the government. Tourists come to see Amsterdam's historic attractions and collections of great art. They admire the city’s scenic canals, bridges, and stately old houses. Amsterdam is also famous for its atmosphere of freedom and tolerance.

City and port, western Netherlands, located on the IJsselmeer and connected to the North Sea. It is the capital and the principal commercial and financial center of the Netherlands. To the scores of tourists who visit each year, Amsterdam is known for its historical attractions, for its collections of great art, and for the distinctive color and flavor of its old sections, which have been so well preserved. However, visitors to the city also see a crowded metropolis beset by environmental pollution, traffic congestion, and housing shortages. It is easy to describe Amsterdam, which is more than 700 years old, as a living museum of a bygone age and to praise the eternal beauty of the centuries-old canals, the ancient patrician houses, and the atmosphere of freedom and tolerance, but the modern city is still working out solutions to the pressing urban problems that confront it. Amsterdam is the nominal capital of the Netherlands but not the seat of government, which is The Hague. The royal family, for example, is only occasionally in residence at the Royal Palace, on the square known as the Dam, in Amsterdam.

- Always monolingual
- Seems to be trivial
  - Use cosine or your favorite similarity function
  - Works well for extraction-based methods, but is problematic in the presence of rewriting
It is hard!

- Insertions, deletions, reordering
- Weak similarity function

| (A) | Petersburg served as the capital of Russia for 200 years.  
|     | For two centuries Petersburg was the capital of the Russian Empire. |

| (B) | The city is also the country’s leading port and center of commerce.  
|     | And yet, as with so much of the city, the port facilities are old and inefficient. |

---

**Sentence alignment in MT**

- Alignment task: Given bitext, identify units which are translations of each other.
- Units: paragraphs, sentences, phrases, words.
- Usage: first step for full translation (Brown et al), lexicography (Dagan & Church, Fung & McKeown), aid for human translators (Shemtov), multi-lingual IR.

---

**Patterns of Mapping**

---

**Length-based Alignment**

- Matching Predicate: Long sentences will be translated as long sentences, short sentences translated as short sentences
- Method: Dynamic programming
**Length-based Alignment**

Let $D(i, j)$ be the lowest cost alignment between sentences $s_1, \ldots, s_i$ and $t_1, \ldots, t_j$.

Base: $D(0, 0) = 0$.

$$D(i, j) = \min \begin{cases} 
D(i, j - 1) + \text{cost}(0:1 \text{ align } \phi, t_j) \\
D(i - 1, j) + \text{cost}(1:0 \text{ align } s_i, \phi) \\
D(i - 1, j - 1) + \text{cost}(1:1 \text{ align } s_i, t_j) \\
D(i - 1, j - 2) + \text{cost}(1:2 \text{ align } s_i, t_{j-1}, t_j) \\
D(i - 2, j - 1) + \text{cost}(2:1 \text{ align } s_{i-1}, s_i, t_j) \\
D(i - 2, j - 2) + \text{cost}(2:2 \text{ align } s_{i-1}, s_i, t_{j-1}, t_j) 
\end{cases}$$

**Design Choices in Alignment**

Determined by a Corpus Type

- Matching predicate
- Search strategy

**Corpus Type**

- Language Proximity (Monolingual vs Bilingual, technical vs lay)
- Content Proximity (comparable vs parallel)
- Matching Granularity (1:1 vs 1:5)

**Matching Predicate**

- Length similarity. (Gale & Church, Brown et al)
- Lexical similarity:
  - Bilingual dictionary (Wu)
  - Words with the same distribution. (Kay & Roscheisen, Fung & McKeown)
  - Cognates (Simard et al, Church, Melamed)
Methods for Overall Alignment

- Dynamic programming
- Methods based on Computational Geometry
- Signal processing Methods

Computational Geometry Methods

(Melamed, 1997) Assumption: Distribution of “true points of correspondence (TPC)” satisfies certain geometric properties

- Generate all the matching points satisfying the matching predicate (over-generation)
- Find a subset of matching points that satisfies a pattern of TPC:
  - Linearity
  - Injectivity
  - Low variance of slope

Various heuristics are used to minimize the search space

Supervised Approaches

Training Data Generation via Alignment → Feature Selection → Classification

Rewriting
**Abstraction: Shallow Features**

Shallow Features:
- Location (in the newspaper genre, the first paragraph is a summary)
- Presence of cue words (e.g., “in conclusion”)
- Sentence length
- Number of highly weighted words in a sentence

**Zipf Distribution**

The product of the frequency of words $(f)$ and their rank $(r)$ is approximately constant: $f \times R = C$ (where $C$ is around $N/10$)

**Assigning Weights**

- Raw frequencies (typically with the list of stop-words)
- TF*IDF – a way to deal with the problem of the Zipf distribution
  - TF - Term frequency
  - IDF - Inverse term frequency

(from van Rijsbergen, 1979) The most frequent words are not the most descriptive
**TF-IDF**

\[ w_{ik} = Tf_{ik} \times \log(N/n_k) \]

- \( w_{ik} \) — Term \( k \) in document \( D_i \)
- \( Tf_{ik} \) — Frequency of term \( k \) in document \( D_i \)
- \( N \) — total number of documents in the collection \( C \)
- \( n_k \) — total number of documents in the collection \( C \) that contain \( T_k \)

**Classiﬁcation for Content Selection**

(Kupiec, Pedersen, Chen, 1995)

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

\[
= \frac{\prod_{j=1}^{k} P(F_j|s \in S)P(s \in S)}{\prod_{j=1}^{k} P(F_j)}
\]

- \( P(s \in S|F_1, \ldots, F_k) \) — Probability that \( s \) is in summary \( S \), given the features
- \( P(s \in S) \) — Probability that \( s \) is in summary \( S \)
- \( P(F_j|s \in S) \) — Probability of feature-value pair for summary sents
- \( P(F_j) \) — Probability of feature-value pair

**Abstraction: Structural Approaches**

“Deep Features”

- Rhetorical structure based
  - RST (Marcu, 2000)
  - Domain-dependent argumentative structure (Teufel&Moens, 2000)
- Content-based (Barzilay&Lee, 2003)

**Supervised Approaches**

- Alignment (trivial for extraction, hard for generation)
- Abstraction
- Classification (standard classiers — Naive Bayes, SVM, maximum entropy, Boostexter)
- Rewriting (Optional)
Rewriting: Title Generation

(see handout C)

- Select candidate content words
- Organize them into English phrases

The score of the title $w_1 \ldots w_n$ is computed as:

$$\prod_{i=1}^{k} p(w_i|w_i \in \text{doc}) \cdot p_{LM}(w_1 \ldots w_k)$$

Semantic-based Summarization

Assumption: In a limited domain, we know “what is important” (Radev&McKeown, 1995, Elhadad&McKeown, 2001)

- Use an information extraction system to select “important information”
- Use a semantics-to-text generation system to generate a new text

Rewriting: Sentence Compression

(see handout D)

- Noisy-Channel Model for Compression (start with short sentence)
- Source Model $P(s)$ gives the likelihood that $s$ is generated as an “original short string”
- Channel Model $P(t|s)$ gives the likelihood that when the short strings $s$ is expanded, the result is the long string $t$

$$\arg\max_s P(s|t) = \arg\max_s \frac{P(s)P(t|s)}{P(t)} = \arg\max_s P(s)P(t|s)$$

Summary Example

Summary:

By multivariate analysis, predictors of sotalol efficacy included age < 60 years, higher left ventricular ejection fraction, and absence of hypertension (1). Neither prior electrical cardioversion nor coronary artery disease predicted sotalol efficacy (1).

(2) identified left atrial size as an independent predictor of conversion, but (3) did not. Age and NYHA class were found not to predict conversion (3). Age, sex, and heart rate were not associated with conversion (2).

In a multivariate analysis, the mode of cardioversion was not associated with the recurrence of atrial fibrillation (4). In both univariate and multivariate analysis, coronary artery disease was an independent predictor of recurrence of atrial fibrillation (4).
New acute myocardial infarction or death was associated with ST-segment depression (OR 2.00, 95% CI 1.20 to 3.40; P = .008), prior angina (OR 2.70, 95% CI 1.34 to 5.57; P = .001), and age > 65 years (OR 1.64, 95% CI 1.00 to 2.70; P = .01).
**Summarization Evaluation: Intrinsic Measures**

Compare against “gold” standard

- Precision/Recall or their weighted version are used
- As a baseline, a “lead” summary is used
- Human agreement is computed using Kappa
- When evaluation results matter, it is done manually (DUC competition)
  - Provides large collection of human-generated summaries
  - Outputs are evaluated manually

---

**Summarization Evaluation: Extrinsic Measures**

Evaluation in the context of a task

- Categorization Task
  - Given a set of summaries or texts, do readers categorize them in the same bins?
- Ad Hoc IR task
  - Given a set of summaries created by a certain query, do readers judge them as relevant?