

Automatic Summarization

Berlin Chen

Department of Computer Science & Information Engineering
National Taiwan Normal University

Outline

- Introduction
- Spectrum of Summarization Research
- Extractive Summarization Methods
- Evaluation Metrics
- Applications
- Conclusions

This talk gives only a partial picture of automatic summarization research, biased and subject to the presenter's expertise.

For more detailed reviews on the developments of automatic summarization, please also refer to, among others,

1 Nenkova and McKeown, "Automatic Summarization," Foundations and Trends in Information Retrieval, 2011

2. Torres-Moreno, "Automatic Text Summarization," Wiley-ISTE, 2014.

Introduction: Information Overload

- Content Creation vs. Content Management
 - *Too much information kills information!*

Text

News articles

Scientific Articles

Books

Websites

Emails

Social Media Streams

© Original Artist
Reproduction right by permission from
www.CartoonStock.com

"It takes 11 ms more time. Each of these beeps is a heartbeat, and an incoming e-mail."

twitter

f

Speech & Multimedia

Phone Conversation

Meeting

Lecture

Classroom

Talk Shows

Chat

Broadcast News

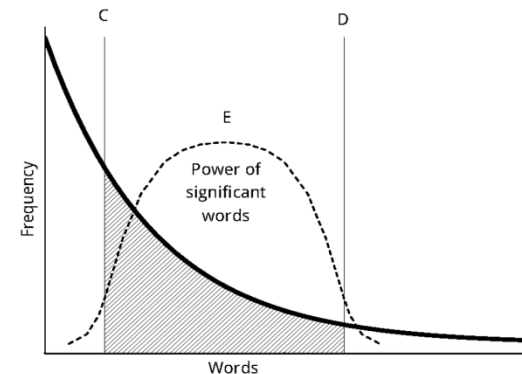
Radio News

Today a person is subjected to more new information in a day than a person in the middle ages in his entire life!

- Automatic summarization figures prominently in dealing with the information overload problem
 - Facilitate people to browse multimedia documents and distill their themes both efficiently and effectively

Introduction: Seminal Work (1/2)

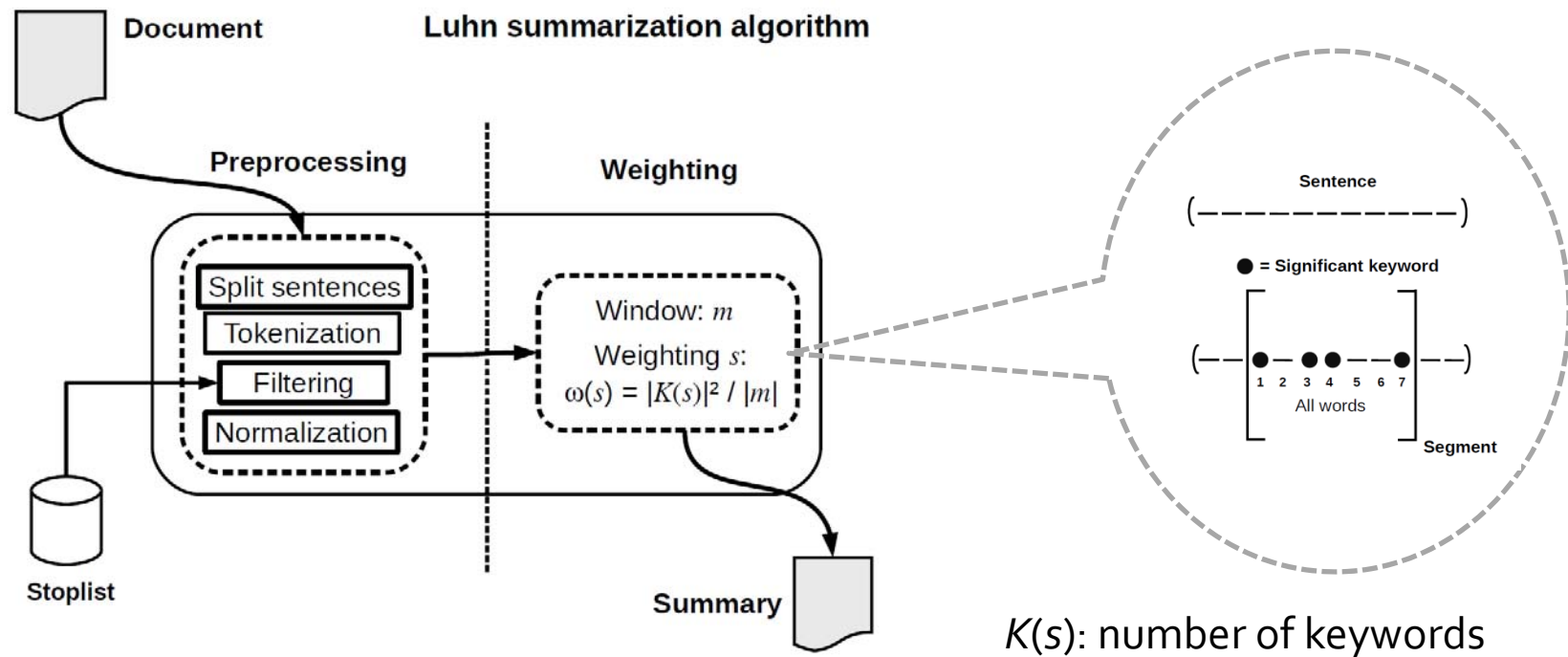
- Developments in automatic summarization date back to the late 1950s (Luhn, 1958) and have continued to be the focus of much research
 - *Luhn put forward a simple idea that shaped much of later research*
 - *Namely, some words in a document are descriptive of its content, and the sentences that convey the most important information in the document are the ones that contain many such descriptive words close to each other*
(Mean what? **frequency, proximity and burstiness/structure** of different levels of lexical/semantic/syntactic units?)
- Nowadays, this research realm is extended to cover a wider range of tasks, including multi-document, multilingual and multimedia (e.g., speech) summarization



1., H. E Luhn , " The automatic creation of literature abstracts," IBM Journal of Research and Development, 1958
2. Nenkova and McKeown, "Automatic Summarization," Foundations and Trends in Information Retrieval, 2011

Introduction: Seminal Work (2/2)

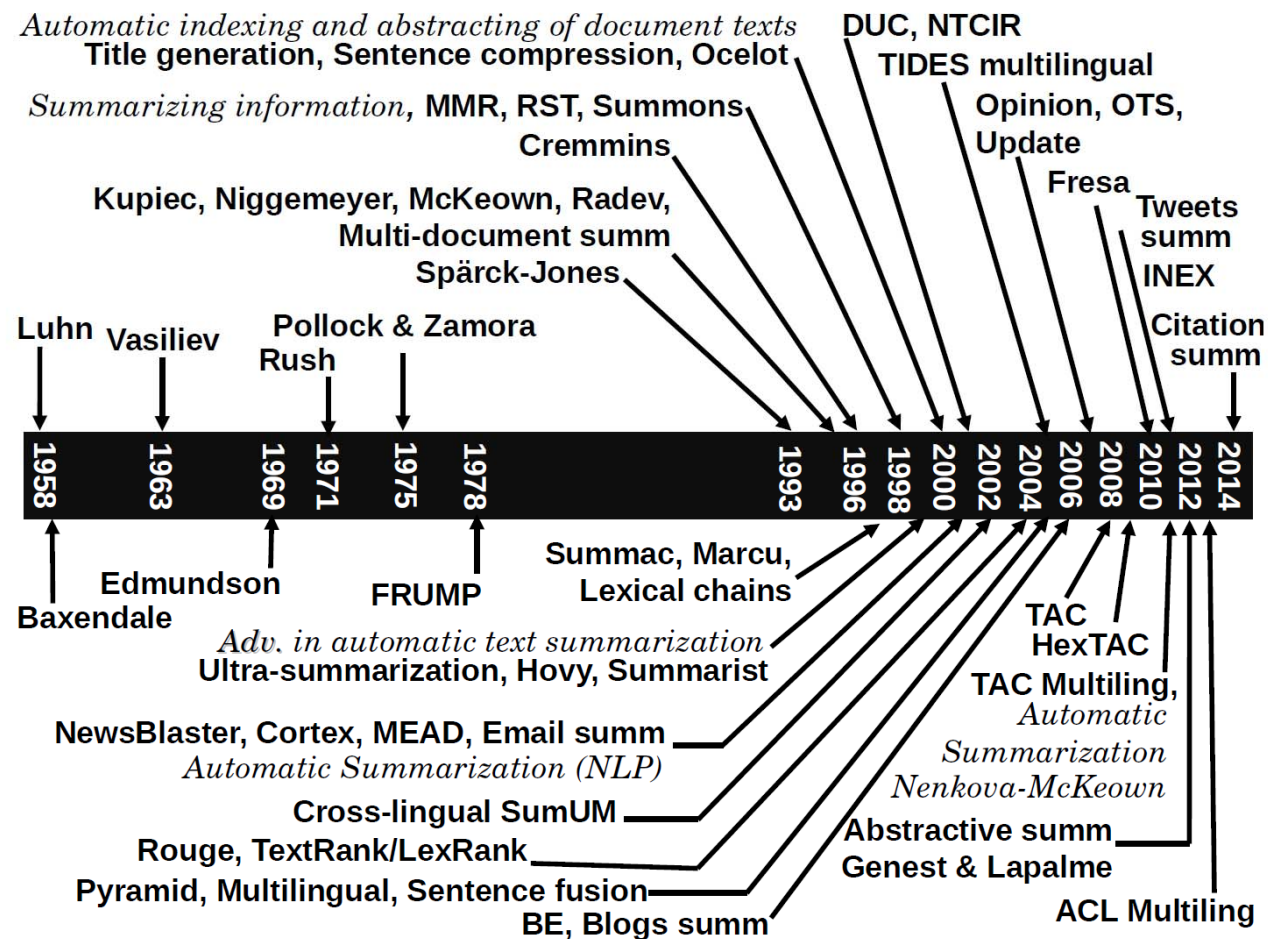
- Schematic Description of Luhn's Method



- A window of the size m words is defined as the longest segment in a sentence between two keywords

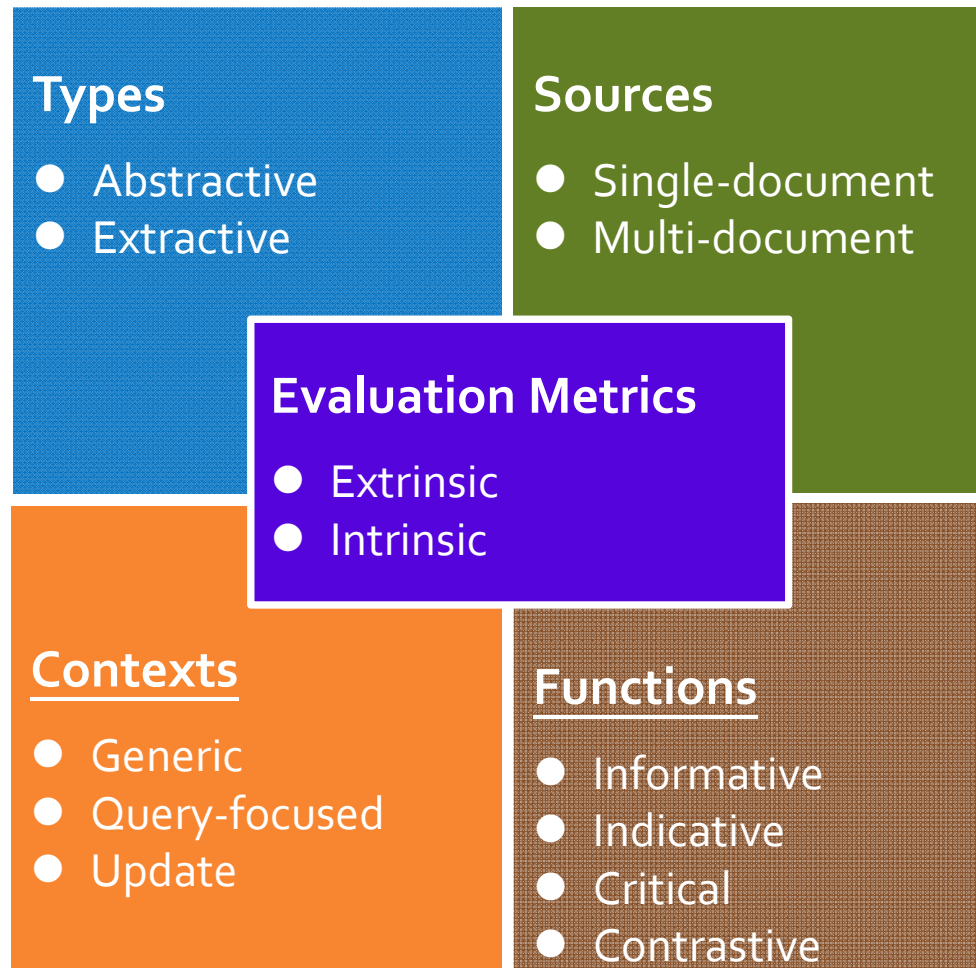
Historical Evolution

- Timeline of Summarization Research



This figure is adopted from Torres-Moreno, "Automatic Text Summarization," Wiley-ISTE, 2014.

Spectrum of Summarization Research



1. Sources

- A summary can be produced from ***a single document*** (single- document summarization) or ***multiple documents*** (multi-document summarization)
- For the latter case, ***information redundancy*** and ***event causality*** (or ordering) are the two main issues concerned, since the information is gathered from several documents

2. Functions (1/2)

- An ***informative*** summary is a condensed presentation which reflects the overall content (ideas/facts) of the original document(s)
 - It usually acts as a surrogate for (can be read in replace of) the original document(s)
- An ***indicative*** summary may provide characteristics such as topics, lengths and writing styles of the original document(s) but does not convey the detailed information of the original document(s)

2. Functions (2/2)

- A **critical** summary provides judgement (either positive and negative) on the input document(s)
 - E.g., *Article review is a critical summarization of another article. Review is basically a productive analysis of an article by summarizing the main points discussed in the topic and by classifying, comparing and assessing the original article critically*
- A **contrastive** summary is formed by automatically extracting and summarizing the multiple contrastive viewpoints implicitly expressed in the opinionated document(s), allowing for digestion and comparison of different viewpoints

1. <http://bigtopmagazine.com/how-to-write-a-unique-article-review-from-scratch.asp>
2. M. J. Paul et al., "Summarizing contrastive viewpoints in opinionated text," EMNLP 2010

2. Functions:

Contrastive Summarization- An Example (1/2)

- 2010 U.S. Healthcare Legislation

- 948 verbatim responses from Gallup opinion phone survey
- 45% for, 48% against (March 2010)

For: “because a lot of people **can't afford it** [insurance]; 45,000 people **die** each year because of lack of healthcare.”

Against: “everybody should have their own healthcare, and if you **can't afford it**, you should just **die**.”

Different viewpoints

Same issue

2. Functions:

Contrastive Summarization- An Example (2/2)

- Make the viewpoint summaries more comparable

For the healthcare bill	Against the healthcare bill
<ul style="list-style-type: none">• i favor healthcare for who needs it, mostly old people who don't have healthcare. the government should help the people when they are old. they should have that kind of healthcare.• i just think something has to be done, the price of health is going up.• [i] pay for private insurance.• bring down cost.	<ul style="list-style-type: none">• i think we can't be responsible for other people's healthcare.• doesn't address things that need to be done, addresses things that don't need to be done.• it's going to increase the cost to those insured.• i believe we can't afford it.• way too expensive, too intrusive, too much government control.

3. Contexts

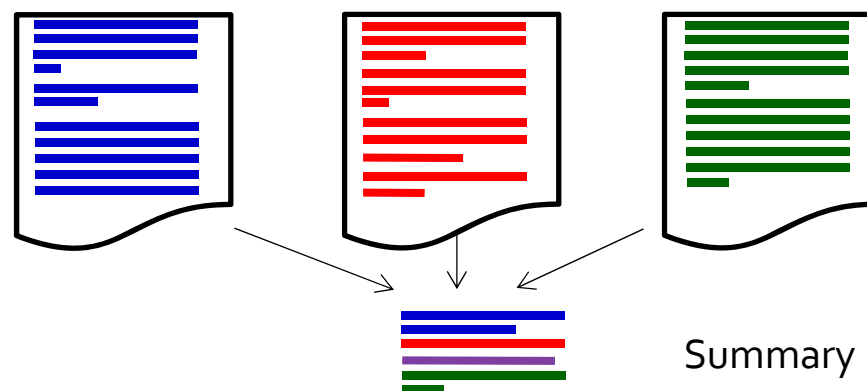
- A summary can be either ***generic, query-oriented, or updated***
 - In ***generic summarization***, each summary provides a general point of view of the original document(s) without regarding to any specific information need
 - ***Query-oriented summarization***, by contrast, is primarily concerned with producing a concise summary that is related to some specific topic (or information need)
 - ***Update summaries only show important new information and avoid repeating information when users are familiar with a particular topic (it is presumed that they have already read documents and their summaries relating to this topic)***

4. Types (1/2)

- Abstractive Summarization
 - Generate a fluent and concise abstract (rewrite a short series of sentences), reflecting the most important information of an original document or a set of documents
 - Require highly sophisticated ***natural language processing (NLP)*** techniques, including semantic representation and inference, as well as natural language generation
 - *Writing a concise and fluent summary requires the capability to reorganize, modify and merge information expressed in different sentences in the input. Full interpretation of documents and generation of abstracts is often difficult for people, and is certainly beyond the state of the art for automatic summarization*

4. Types (2/2)

- Extractive Summarization
 - Select a set of salient sentences from an original document or a set of documents (according to a predetermined target summarization ratio), and concatenate them to form a summary
 - Typical operations includes **sentence selection** (ranking) [mandatory], **sentence ordering** [mandatory], **sentence fusion** [optional], **sentence revision** [optional] and **sentence compression** [optional], etc.



5. Methodology for Evaluation (1/2)

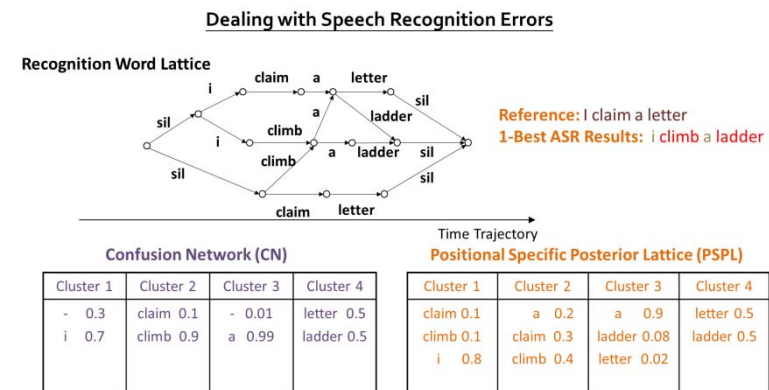
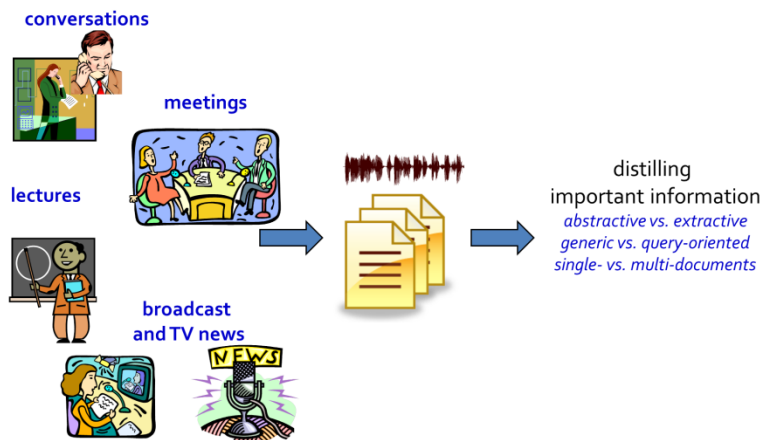
- Research on automatic summarization has frequently been criticized for lacking ideal (unanimous) “**gold-standard**” summaries when evaluating the performance of given automatically generated summaries
- However, current evaluation approaches can generally be classified as either *extrinsic* or *intrinsic*

5. Methodology for Evaluation (2/2)

- **Extrinsic evaluation** measures the impact of the summary on the performance of downstream applications (tasks)
 - Such as information retrieval, document classification, essay scoring, among others
 - *Could be time-consuming, expensive and require a considerable amount of careful planning*
- **Intrinsic evaluation** examines how well a summarizer performs in relation to human experts
 - Usually, it can be done by comparing the automatic summaries output from the summarizer to those provided by human experts (in terms of **objective** and **subjective** quality)

Speech Summarization

- Speech Summarization vs. Text Summarization
 - Speech summarization inevitably suffers from the problems of **recognition errors** and **incorrect sentence boundaries** when using ASR techniques to transcribe the spoken documents into text forms
 - On the other hand, speech summarization also presents information cues that are peculiar to it and do not exist for text summarization, such as information cues about **prosodies/acoustics** and **emotions/speakers**, which can potentially help in determining the important parts or implicit structures of spoken documents

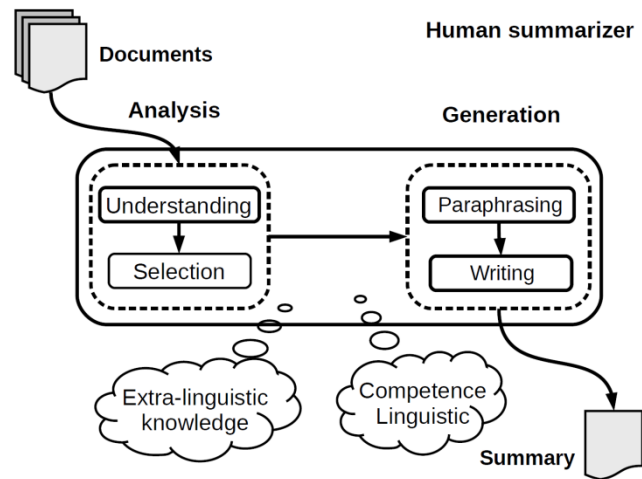


1. Liu and Hakkani-Tur, "Speech summarization," in Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, 2011
2. Lin et al., "Leveraging Kullback-Leibler divergence measures and information-rich cues for speech summarization," IEEE Transactions on Audio, Speech and Language Processing, 2011

How Humans Produce Summaries?

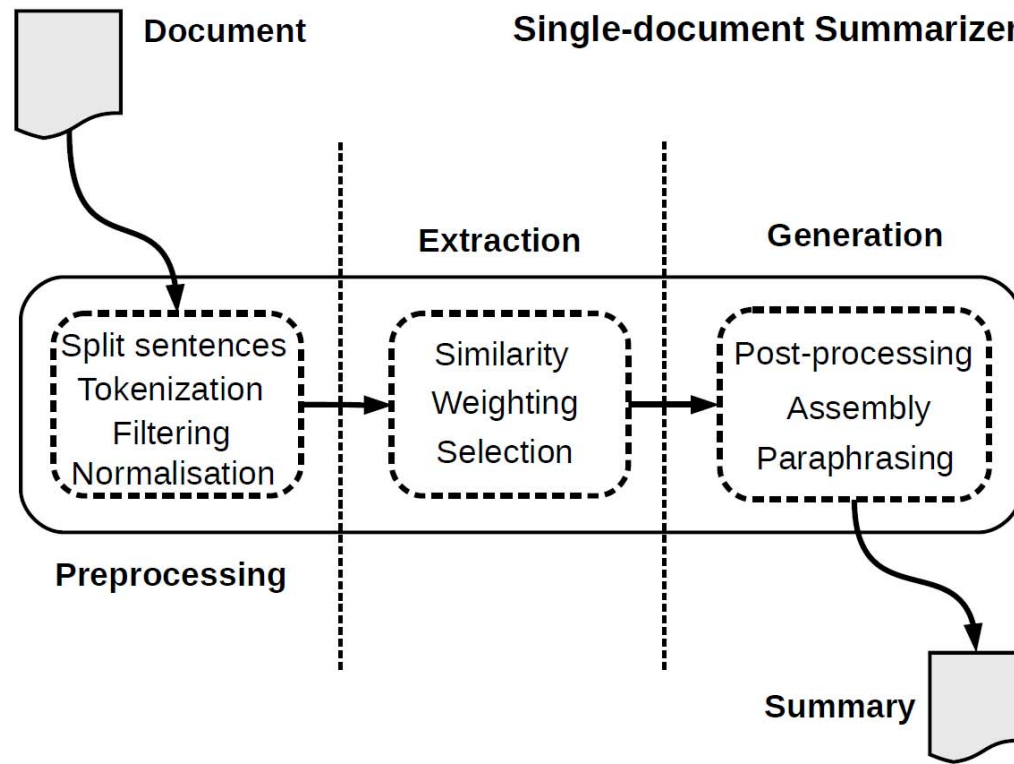
- The human production of summaries may involve two phases:
 - First, understand and interpret the source text
 - Then, write a concise and shortened version of it

*Both require **linguistic and extra-linguistic skills and (world) knowledge on the part of the summarizer***



- *Is it worth modeling and trying to replicate the abstracting process of humans?*

More on Extractive Summarization



This figure is adopted from Torres-Moreno, "Automatic Text Summarization," Wiley-ISTE, 2014.

Considerations when Conducting Extractive Summarization

- A sentence to be selected as part of a summary may be considered from the following three factors
 - 1) ***Salience***—the importance of the sentence itself, which is usually evident by its structure, location or word-usage information, and many more
 - 2) ***Relevance***—the more relevant a sentence to the input document(s) or the other sentences in the document(s), the more likely it should be included in the summary
 - 3) ***Redundancy***—the information carried by the sentence and that of the already selected summary sentences should cover different topics or concepts of the document(s)

Categorization of Extractive Summarization Methods

- The wide array of ***extractive summarization*** methods may roughly fall into three main categories:
 - Approaches simply based on ***sentence structure or position information, word-level statistics***
 - LEAD
 - Approaches based on ***unsupervised machine-learning***
 - Vector-based methods
 - VSM, LSA, MMR
 - Graph-based methods
 - TextRank, LexRank, MRW
 - Combinatorial optimization-based methods
 - ILP, Submodular
 - Language modeling methods, word/sentence embeddings
 - Approaches based on ***supervised machine-learning***
 - GMM, BC, SVM, CRF, DNN/RNN/LSTM

LEAD

- The **LEAD** method condenses an input document using only the first portion of document (the lead; e.g., the first several sentences) until the target length of the summary is reached
- **LEAD** was known to be effective than other methods (at the time of its day) for document summarization of newspapers in lower summarization ratio

Method Using Simple Word-level Statistics

- Word Probability (SumBASIC)
 - Calculate the word probability for **content words** in a document to be summarized

$$P(w) = \frac{c(w)}{N}$$

- Determine the importance of a sentence based on the average probabilities of **content words** involved in the sentence

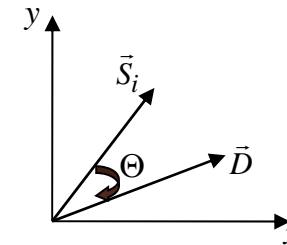
$$\text{Weight}(S) = \frac{\sum_{w \in S} P(w)}{|\{w \mid w \in S\}|}$$

- After the best sentence is selected, the probability of each word that appears in the chosen sentence is adjusted, set to a smaller value (?)
- Then select another best-scoring sentences from the rest until the desired summary length is achieved

Vector Space Model (VSM)

- Represent sentences and the document to be summarized as vectors using statistical weighting such as the product of **Term Frequency** (TF) and the **Inverse Document Frequency**
- Sentences are ranked based on their similarity to the document

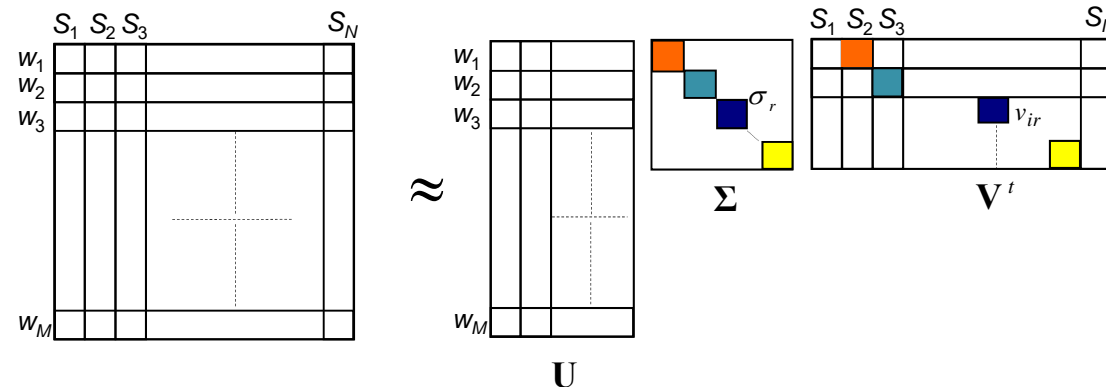
$$\text{SIM}(S, D) = \frac{\vec{S} \cdot \vec{D}}{|\vec{S}| |\vec{D}|}$$



- To summarize more important and different concepts in a document
 - The terms occurring in the sentence with the highest relevance score $\text{SIM}(S, D)$ are removed from the document
 - The document vector is then reconstructed and the ranking of the rest of the sentences is performed accordingly

Latent Semantic Analysis (LSA)

- Construct a “term-sentence” matrix for a given document
- Perform SVD on the “term-sentence” matrix
 - The **right singular vectors** with larger singular values represent the dimensions of the more important latent semantic concepts in the document
 - Represent each sentence of a document as a vector in the latent semantic space
- Sentences with the largest index (element) values in each of the top L right singular vectors are included in the summary



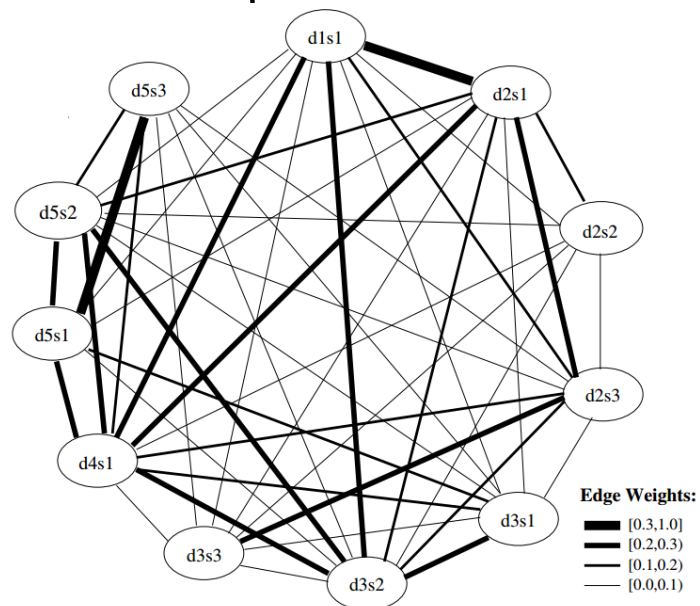
Maximum Marginal Relevance (MMR)

- Perform sentence selection iteratively with the criteria of topic relevance and coverage
- A summary sentence is selected according to
 - Whether it is more similar to the whole document than the other sentences (**Relevance**)
 - Whether it is less similar to the set of sentences selected so far than the other sentences (**Redundancy**)

$$S_{MMR} = \arg \max_{S_i} \left[\beta \cdot Sim(S_i, D) - (1 - \beta) \cdot \max_{S' \in \text{Summ}} Sim(S_i, S') \right]$$

Graph-based Methods (1/2)

- Graph-based methods, such as *TextRank* and *LexRank*, conceptualize the document to be summarized as a network of sentences
 - Where each node represents a sentence and the associated weight of each link represents the lexical or topical similarity relationship between a pair of nodes



TextRank

1. Mihalcea et al., "TextRank: Bringing order into texts," EMNLP 2004
2. Günes et al., "LexRank: graph-based lexical centrality as salience in text summarization." Journal of Artificial Intelligence Research, 2004

Graph-based Methods (2/2)

- After constructing the conceptualized network, a graph-based centrality algorithm is then applied to obtain an importance score for each sentence
- The network actually can be viewed as a Markov chain in which the states are the sentences and the corresponding state transition distribution is given by a similarity matrix \mathbf{W}
- Then, the importance of each sentence can be derived by the following equation

$$\text{WS}(v_i) = (1 - \alpha) + \alpha \times \sum_{v_j \in \text{In}(v_i)} \frac{w_{ji}}{\sum_{v_k \in \text{Out}(v_j)} w_{jk}} \cdot \text{WS}(v_j)$$

Prestige Score

$$w_{ij} = w_{ji} = \text{SIM}(S_i, S_j) = \frac{\sum_{w \in S_i, S_j} c(w, S_i) + c(w, S_j)}{\log(|S_i|) + \log(|S_j|)}$$

TextRank

Integer Linear Programming (ILP) (1/2)

- The ILP-based summarization methods have gained considerable attention since they, to some extent, solve the sub-optimal (greedy) summary sentences selection problem
 - Summary Selection vs. Sentence Selection
- ILP is developed for the constrained optimization problem, where both the cost function and constraint are linear in a set of integer variables

Integer Linear Programming (ILP) (2/2)

- The extractive summarization task is formulated as a maximum convergence problem which is subjected to a set of defined objective functions and summary-length constraint

$$\text{maximize } \sum_i \alpha_i \text{Relevance}(i) - \sum_{i < j} \beta_{ij} \text{Redundancy}(i, j)$$

$$\text{subject to } (1) \alpha_i, \beta_{ij} \in 0, 1$$

$$(2) \sum_i \alpha_i l(i) \leq K$$

$$(3) \beta_{ij} - \alpha_i \leq 0$$

$$(4) \beta_{ij} - \alpha_j \leq 0$$

$$(5) \alpha_i + \alpha_j - \beta_{ij} \leq 1$$

- $\text{Relevance}(i)$ is the relevance degree of sentence S_i to the entire document
- $\text{Redundancy}(i, j)$ is defined as the similarity between sentence pairs S_i and S_j

Language Modeling (LM) Methods (1/3)

- Document-Likelihood Measure (DLM)

- Each sentence S of a document D is treated as a probabilistic generative model for generating the document
- The higher the probability $P(D/S)$, the more representative S is likely to be for D

$$P(S | D) = \frac{P(D|S)P(S)}{P(D)} \approx P(D | S)$$

Sentence Modeling $P(D | S) \approx \prod_{w \in D} P(w | S)^{c(w,D)}$

Maximum Likelihood
Estimation (MLE)

$$\frac{c(w, S)}{|S|}$$

- Kullback-Leibler Divergence Measure (KLM)

- Rank a sentence according its **model distance** to the document

$$KL(D || S) = \sum_{w \in V} P(w | D) \log \frac{P(w | D)}{P(w | S)}$$

- KLM is reduced to DLM when $P(w|D)$ is computed with MLE

Language Modeling (LM) Methods (2/3)

- Extension 1: ***Sentence Clarity Measure***

- Quantify the thematic specificity of each candidate summary sentence

$$Clarity(S) \stackrel{def}{=} CE(B || S) - H(S)$$

- Where

$$CE(B || S) = -\sum_{w \in V} P(w/B) \log P(w/S)$$

- $CE(B||S)$ is the cross entropy between the background unigram model $P(w/B)$ and the sentence model $P(w/S)$
- It is hypothesized that the higher the cross entropy (or the farther the sentence model away from the background model), the more thematic information the sentence S is to convey

$$-H(S) = \sum_{w \in V} P(w|S) \log P(w|S)$$

- The higher the negative entropy $-H(S)$, the more concentrative the word usage of the sentence S is, revealing that S concentrates more on some important aspect of the document

- Sentence Ranking

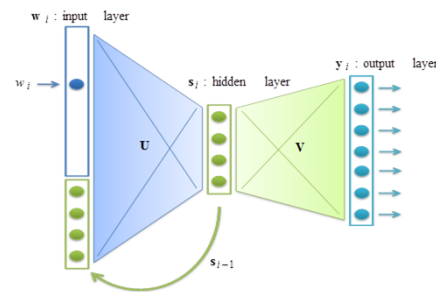
$$-KL(D || S) + Clarity(S)$$

Language Modeling (LM) Methods (3/3)

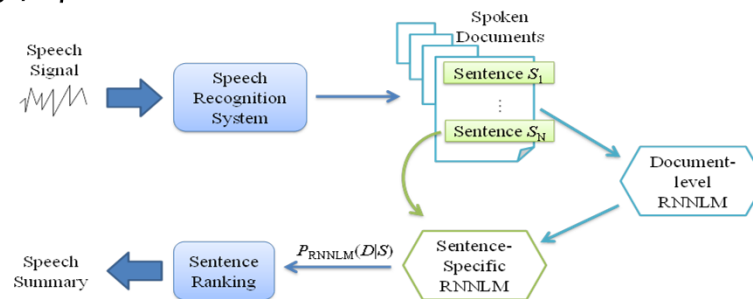
- Extension 2: **Recurrent Neural Networks (RNN)**

- Leverage RNN for sentence modeling

$$P_{\text{RNNLM}}(D | S) = \prod_{i=1}^L P_{\text{RNNLM}}(w_i | w_1, \dots, w_{i-1}, S)$$



- Schematic Depiction
E.g., speech summarization



Input:

H : Number of Hidden Layer Neurons

$$\mathbf{D} = \{D_1, \dots, D_m, \dots, D_M\}$$

$$D_m = \{S_1^{D_m}, \dots, S_j^{D_m}, \dots, S_{|D_m|}^{D_m}\}$$

Model Training & Important Sentence Ranking:

- 1: **for** D_1 to D_M **do**
- 2: document-level RNNLM model training
- 3: $\mathcal{L}(\mathbf{U}_m, \mathbf{V}_m) = \sum_{i=1}^{|D_m|} \log(y_i)$
- 4: **for** $S_1^{D_m}$ to $S_{|D_m|}^{D_m}$ **do**
- 5: sentence-level RNNLM model training
- 6: $\mathcal{L}(\mathbf{U}_{S_j^{D_m}}, \mathbf{V}_{S_j^{D_m}} | \mathbf{U}_m, \mathbf{V}_m) = \sum_{i=1}^{|S_j^{D_m}|} \log(y_i)$
- 7: **end for**
- 8: **for** $S_1^{D_m}$ to $S_{|D_m|}^{D_m}$ **do**
- 9: calculate document likelihood
- 10: $P(D_m | S_j^{D_m}) = \prod_{i=1}^{|S_j^{D_m}|} P(w_i | w_1, \dots, w_{i-1}, S_j^{D_m})$
- 11: $= \prod_{i=1}^{|S_j^{D_m}|} P(w_i | \mathbf{U}_{S_j^{D_m}}, \mathbf{V}_{S_j^{D_m}}, S_j^{D_m})$
- 12: **end for**
- 13: Sentence selection according to $P(D_m | S_j^{D_m})$
- 14: **end for**

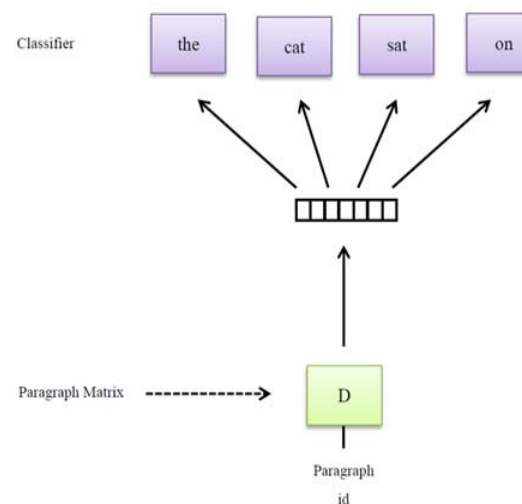
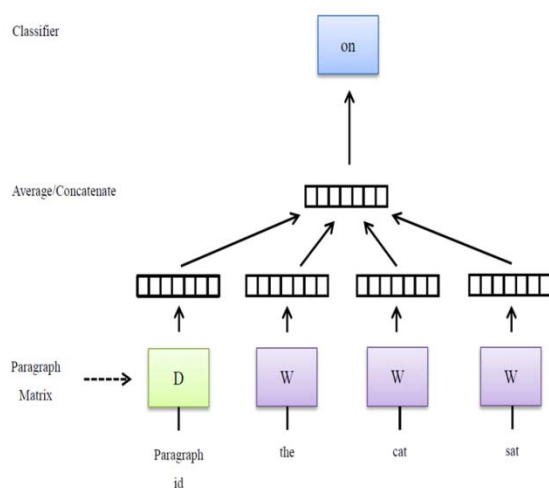
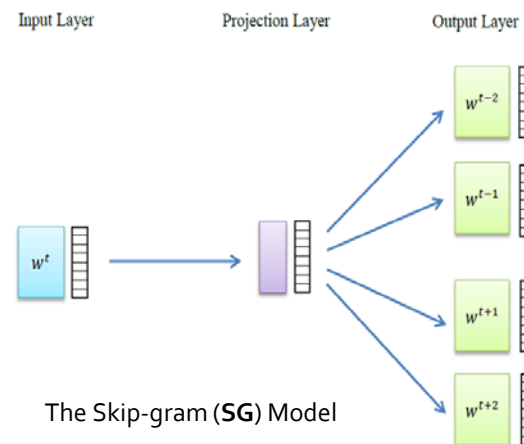
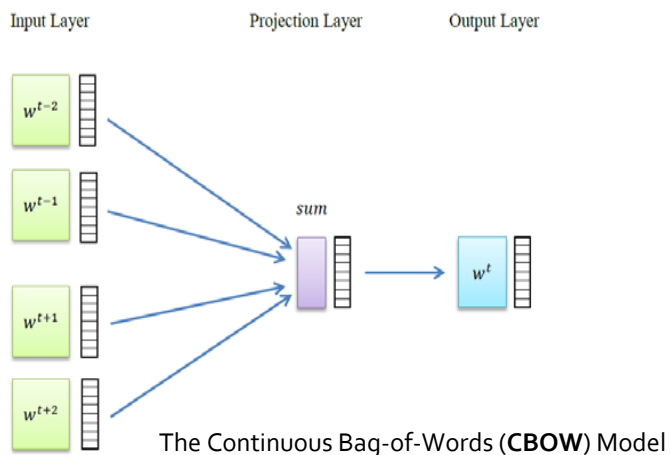
The design of learning curriculum for RNN is of paramount importance here

Representation Learning for Summarization (1/4)

- Word (Sentence) Embeddings (**WE/SE**)
 - Instead of a one-hotspot vector, a word is represented by a real-valued vector with a much smaller size (normally by several hundreds)
 - The syntactic and semantic regularities of words can be encoded in the distributed vector space: the Euclidean distance between two words in the lower-dimensional vector space represents the syntactic or semantic similarity between them
 - E.g., $\text{vector}(\textit{king}) - \text{vector}(\textit{man}) + \text{vector}(\textit{woman})$ results in a vector that is closest to $\text{vector}(\textit{queen})$
 - A common thread of leveraging word embeddings to NLP-related tasks is to represent the document (or query and sentence) by averaging the word embeddings corresponding to the words occurring in the document (or query and sentence)

Representation Learning for Summarization (2/3)

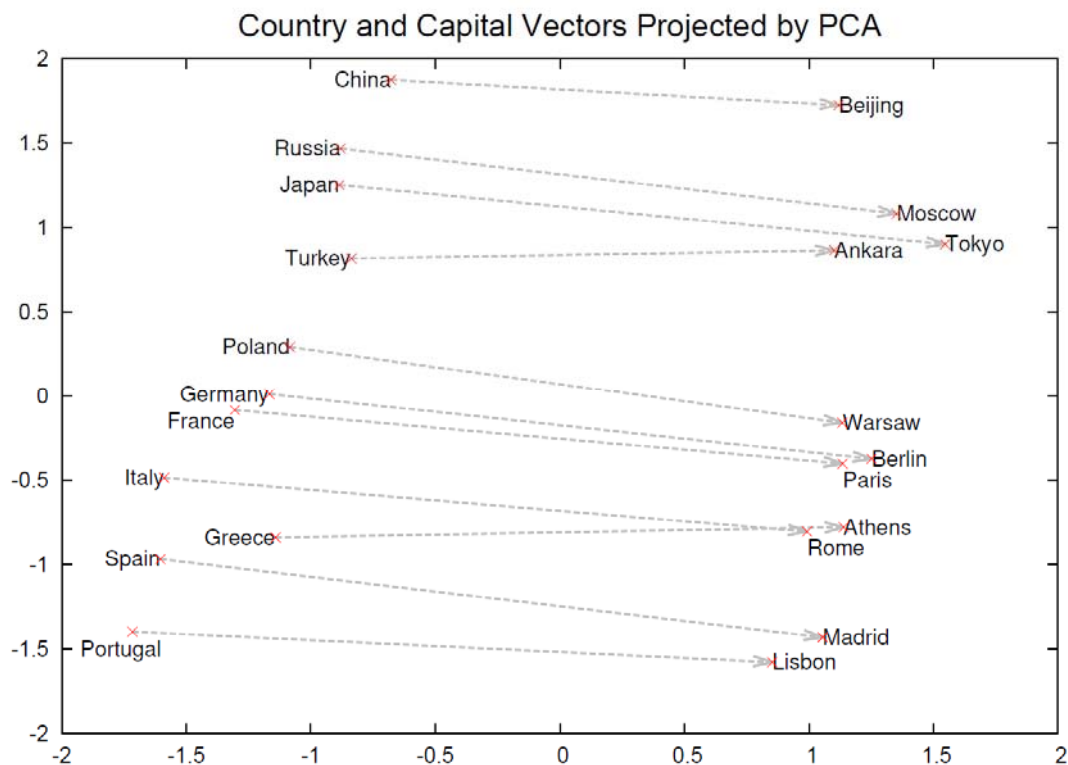
- Some Typical Learning Architectures



The Distributed Memory of Paragraph Vector (PV-DM) Model

The Distributed Bag-of-Words of Paragraph Vector (PV-DBOW) Model

Representation Learning for Summarization (3/4)



Two-dimensional PCA projection of the 1,000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as no any supervised information about what a capital city means was provided during the training .

Representation Learning for Summarization (4/4)

- Incorporate Word/Sentence/Document Embeddings into Extractive Summarization (sentence ranking)

- **Vector Space Model** (Cosine Similarity Measure)

$$\text{SIM}(S, D) = \frac{v_S \cdot v_D}{\|v_S\| \cdot \|v_D\|} \quad \text{where} \quad v_D = \frac{\sum_{w \in D} v_w}{|D|}$$

- **Graph-based Model** (Centrality Measure)

$$\text{WS}(v_i) = (1 - \alpha) + \alpha \times \sum_{v_j \in \text{In}(v_i)} \frac{w_{ji}}{\sum_{v_k \in \text{Out}(v_j)} w_{jk}} \cdot \text{WS}(v_j) \quad \text{where} \quad w_{ij} = w_{ji} = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|}$$

- **Language Model** (Document Likelihood Measure)

$$P(D|S) = \prod_{w_j \in D} \left[\lambda \cdot \sum_{w_i \in S} P_{\text{MLE}}(w_i|S) \cdot P_{\text{WE}}(w_j|w_i) + (1 - \lambda) \cdot P_{\text{MLE}}(w_j|C) \right]^{c(w_j, D)}$$

$$\text{where} \quad P_{\text{WE}}(w_j|w_i) = \frac{\exp(v_i \cdot v_j)}{\sum_{w_k \in V} \exp(v_i \cdot v_k)}$$

Supervised Summarization Methods

- A number of classification-based methods using various kinds of representative (heterogeneous) features also have been investigated
 - In these methods, important sentence selection is usually formulated as a **binary classification** problem
 - A sentence can either be included in a summary or not
- These classification-based methods need a set of training documents along with their corresponding **handcrafted summaries** (or **labeled data**) for training the classifiers (or summarizers)

Support Vector Machines (SVM)

- **SVM** attempts to find an optimal hyper-plane by utilizing a decision function that can correctly separate the *summary* and *non-summary* sentences
- Specifically, SVM to construct a binary summarizer to output the decision score $g(S_i)$ of a sentence S_i
- The posterior probability of a sentence S_i being included in the summary class \mathbf{S} can be approximated by

$$P(S_i \in \mathbf{S} | \mathbf{X}_i) \approx \frac{1}{1 + \exp(\alpha \cdot g(S_i) + \beta)}$$

- In contrast to SVM, **Ranking SVM** seeks to create a more rank- or preference-sensitive ranking function

$$l(S_i) \succ l(S_j) \Leftrightarrow f(S_i) \succ f(S_j)$$

- $l(S_i)$ denotes the preference label of a sentence S_i ; $f(S_i)$ denotes the decision value

Perception

- The decision score of sentence S_i produced by *Perception* is

$$f(S_i) = \boldsymbol{\alpha} \cdot \mathbf{X}_i$$

- That is, the inner product of feature vector of sentence S_i and model parameter $\boldsymbol{\alpha}$
- The model parameter vector of Perceptron can be estimated by maximizing the accumulated *squared score distances* of all the training spoken documents defined as follows

$$F_{\text{Perceptron}}(\boldsymbol{\alpha}) = \frac{1}{2} \cdot \sum_{n=1}^N \sum_{S_R \in \text{Summ}_n} \left(f(S_R) - f(S_n^*) \right)^2$$

- N is total training documents; Summ_n is the reference summary of the n -th training document D_n ; S_R is a summary sentence in Summ_n ; S_n^* is the non-summary sentence of D_n that has the highest decision score

Global Conditional Log-linear Model (GCLM)

- The model parameter vector α of GCLM is estimated by maximizing the following objective function

$$F_{\text{GCLM}}(\alpha) = \sum_{n=1}^N \sum_{S_R \in \text{Summ}_n} \log \frac{\exp(\alpha \cdot \mathbf{X}_R)}{\sum_{S_j \in D_n} \exp(\alpha \cdot \mathbf{X}_j)}$$

- By doing so, the GCLM method will maximize the posterior of the summary sentences of each given training spoken document

More NLP-Intensive Methods

- Yet another school of thought *either relies on existing manually constructed semantic resources (lexical chains, concepts), on coreference tools, or on knowledge about lexical items induced from large collections of unannotated text*
- Most of methods developed along this line of research might tend to be **fragile**, or **difficult** to replicate or extend from constrained domains to more general domains

Lexical Chains (1/2)

- *The lexical chains approach exploits the intuition that topics are expressed using not a single word but instead different related words*
 - *E.g., words "car", "wheel", "seat" and "passenger" indicate a clear topic, even if each of the words is not by itself very frequent*
 - *The methods heavily rely on external hand-crafted resources, such as **WordNet** which lists the different sense of each word, as well as word relationships such as synonymy, antonymy, part-whole and general-specific*
 - *In addition, the lexical chains approach requires some degree of **linguistic preprocessing**, including **part of speech tagging** and **division into topically related segments** of the input to the summarizer*

Lexical Chains (2/2)

- *As an illustration, summarization is conducted by segmenting an input document, identifying lexical chains first within segments and then across segments, identifying and scoring lexical chains, and finally selects one sentence for each of the most highly scored chains*
 - *The strength (score) of a lexical chain, for example, can be computed by its length, defined as the number of words found to be members of the same chain*
- *The core of the problem is how to **construct good lexical chains**, with emphasis on **word sense disambiguation** of words with multiple meaning*
 - *E.g., the word "bank" can mean a financial institution or the land near a river or lake*

Coreference Information

- *Another way of tracking lexically different references to the same semantic entity is the use of coreference resolution*
- ***Coreference resolution** is the process of finding all references to the same entity in a document, regardless of the syntactic form of the reference: full noun phrase or pronoun*
- *However, some initial uses of coreference information exclusively to determine sentence importance for summarization did not lead to substantial improvements in content selection compared to shallower methods*

Rhetorical Structure Theory (1/2)

- *Other research uses analysis of the discourse structure of the input document to produce single document summaries.*
Rhetorical Structure Theory (RST)
 - *It requires the overall structure of a text to be represented by a tree, being one such approach that has been applied to summarization*
- *In RST, the smallest units of text analysis are **elementary discourse units (EDUs)***
 - *They are in most cases sub-sentential clauses*
 - *Adjacent EDUs are combined through rhetorical relations into larger spans*
 - *The larger units recursively participate in relations, yielding a hierarchical tree structure covering the entire text*
 - *The discourse units participating in a relation are assigned **nucleus** or **satellite** status*

Rhetorical Structure Theory (2/2)

- *Properties of the RST tree used in summarization include the nucleus–satellite distinction, notions of salience and the level of an EDU in the tree*

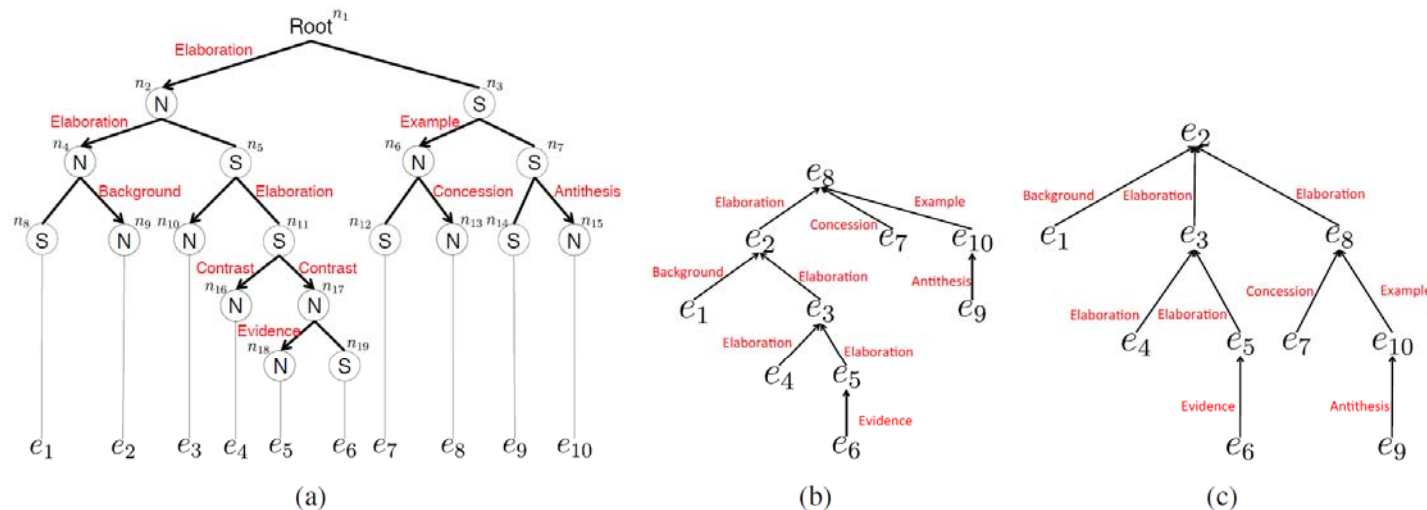
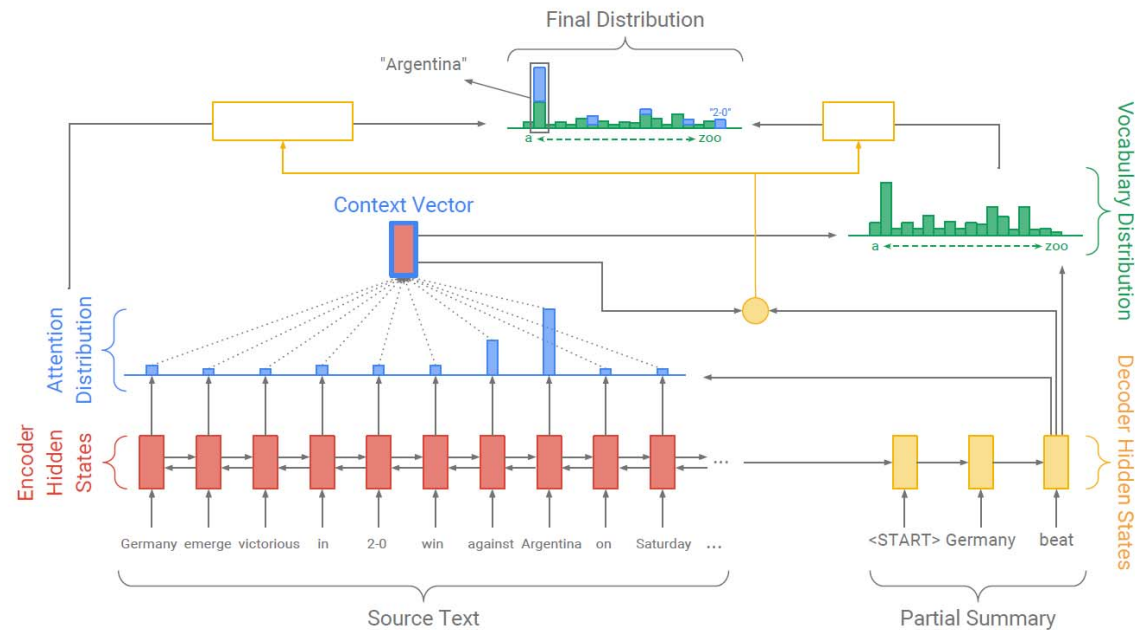


Figure 1: Examples of RST-DT and DEP-DT. e_1, \dots, e_{10} are EDUs. (a) Example of an RST-DT from (Marcu, 1998). n_1, \dots, n_{19} are the non-terminal nodes. (b) Example of the DEP-DT obtained from the incorrect RST-DT that is made by swapping the Nucleus-Satellite relationship of the node n_2 and the node n_3 . (c) The correct DEP-DT obtained from the RST-DT in (a).

Abstractive Summarization with Neural Networks (1/2)

- The recent success of sequence-to-sequence models, in which recurrent neural networks (RNNs) both read and freely generate text, has made abstractive summarization viable



Abstractive Summarization with Neural Networks (2/2)

- Some shortcomings remains to be solved
 - Inaccurately reproducing factual details
 - An inability to deal with out-of-vocabulary (OOV) words
 - Often repeating themselves

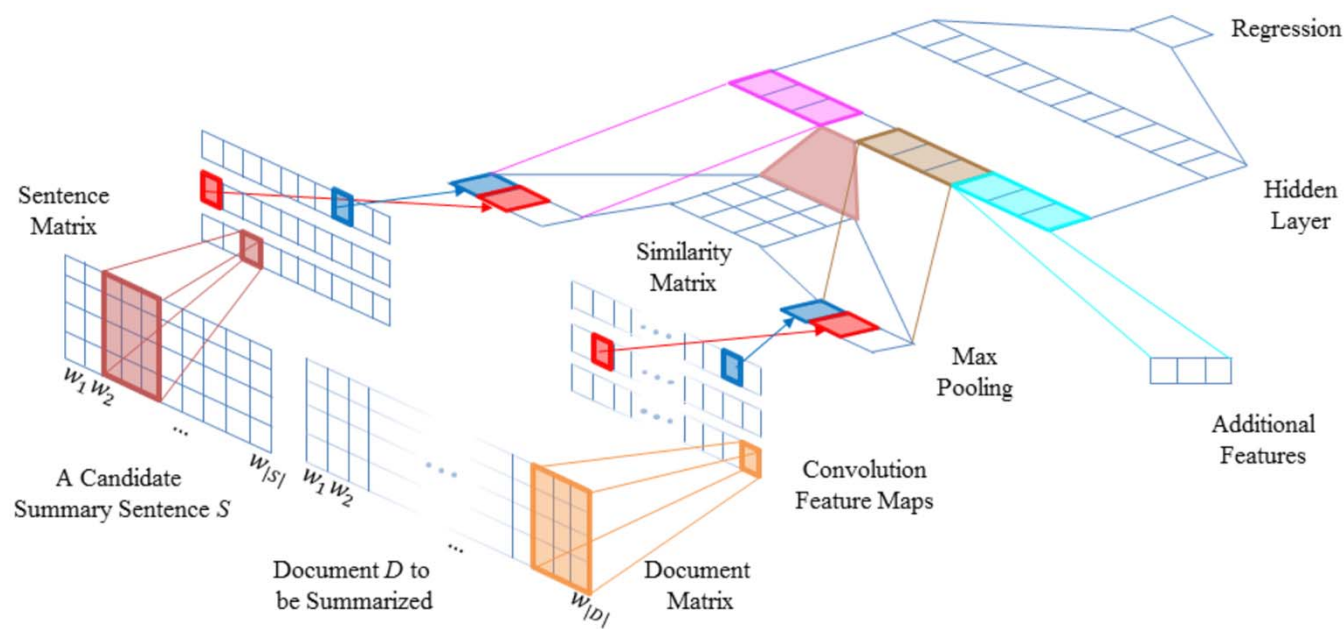
Original Text (truncated): lagos, nigeria (cnn) a day after winning nigeria's presidency, *muhammadu buhari* told cnn's christiane ampanpour that **he plans to aggressively fight corruption that has long plagued nigeria** and go after the root of the nation's unrest. *buhari* said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, **he said his administration is confident it will be able to thwart criminals** and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. *buhari* defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. **the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.**

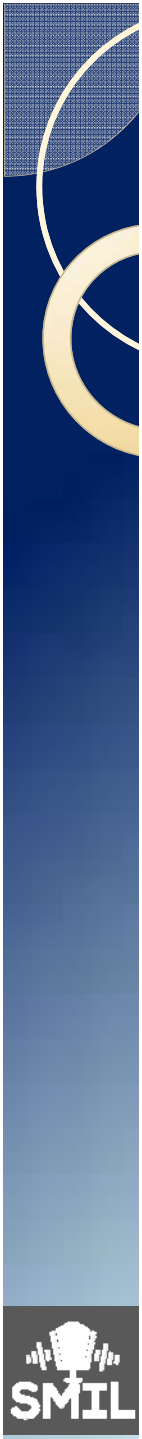
Baseline Seq2Seq + Attention: UNK UNK says his administration is confident it will be able to **destabilize nigeria's economy**. UNK says his administration is confident it will be able to thwart criminals and other **nigerians**. **he says the country has long nigeria and nigeria's economy.**

Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption **in the northeast part of nigeria**. he says he'll "rapidly give attention" to curbing violence **in the northeast part of nigeria**. he says his administration is confident it will be able to thwart criminals.

Pointer-Gen + Coverage: *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

Extractive Summarization with Neural Networks





More on Evaluation Metrics

Intrinsic Evaluations

- Intrinsic evaluations on automatic summaries could be ***objective*** and ***subjective***
- Objective evaluations:
 - Compare the automatic summaries with human-authored (reference) summaries that serve as the gold-standards
 - Less human involvement is usually preferred
- Subjective evaluations:
 - Solicit human judgements on the goodness and utility of automatically generated summaries

Objective Evaluations: Recall and Precision

- **Recall**

- The fraction of the reference (human chosen) summary sentences that are included in the automatic summary

$$\text{Recall} = \frac{|\text{overlap between reference and automatic summary sentences}|}{|\text{reference summary sentences}|}$$

- **Precision**

- The fraction of the reference (human chosen) summary sentences that are included in the automatic summary

$$\text{Precision} = \frac{|\text{overlap between reference and automatic summary sentences}|}{|\text{automatic summary sentences}|}$$

- **F-measure** is the harmonic mean of precision and recall

To avoid susceptibility to bias produced by personal opinions, the above evaluations usually involve multiple reference summaries.

Objective Evaluations: ROUGE (1/2)

- Recall-Oriented Understudy for Gisting Evaluation (**ROUGE**) evaluates the quality of the summarization by counting the number of overlapping units
 - Such as N -grams, longest common subsequences or skip-bigrams, between the automatic summary and a set of reference summaries
 - The ROUGE- N is an N -gram recall measure defined as follows

$$\text{ROUGE-}N = \frac{\sum_{M \in S_{Ref}} \sum_{\text{gram}_N \in M} \text{Count}_{\text{match}}(\text{gram}_N)}{\sum_{M \in S_{Ref}} \sum_{\text{gram}_N \in M} \text{Count}(\text{gram}_N)}$$

Where N denotes the length of the N -gram; M is an individual reference (or manual) summary; S_{Ref} is a set of reference summaries; $\text{Count}_{\text{match}}(\text{gram}_N)$ is the maximum number of N -grams co-occurring in the automatic summary and the reference summary; and $\text{Count}(\text{gram}_N)$ is the number of N -grams in the reference summary

Objective Evaluations: ROUGE (2/2)

- The ROUGE-1 measure evaluates the **informativeness** of automatic summaries
- The ROUGE-2 measure estimates the **fluency** of automatic summaries
- ROUGE-*L* does not reward for fixed-length *N*-grams but instead for a combination of the maximal substrings of words, which works well in general for evaluating both **content** and **grammaticality**

The variants of the ROUGE measure are evaluated by computing the correlation coefficient between ROUGE scores and human judgement scores, while ROUGE-2 performs the best among the ROUGE-*N* variants.

Objective Evaluations: An Example (1/2)

- Broadcast News Summarization
 - Results achieved with **unsupervised methods**

	Text Documents (TD)			Spoken Documents (SD)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
KLM	0.411	0.298	0.371	0.364	0.210	0.307
LEAD	0.310	0.194	0.276	0.255	0.117	0.221
VSM	0.347	0.228	0.290	0.342	0.189	0.287
LSA	0.362	0.233	0.316	0.345	0.201	0.301
MMR	0.368	0.248	0.322	0.366	0.215	0.315
MRW	0.412	0.282	0.358	0.332	0.191	0.291
LexRank	0.413	0.309	0.363	0.305	0.146	0.254
Submodularity	0.414	0.286	0.363	0.332	0.204	0.303
ILP	0.442	0.337	0.401	0.348	0.209	0.306

	Text Documents (TD)			Spoken Documents (SD)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
KLM+Clarity	0.447	0.335	0.393	0.403	0.261	0.354
KLM+RM+Clarity	0.477	0.373	0.426	0.400	0.266	0.354
KLM+WRM+Clarity	0.476	0.367	0.424	0.403	0.263	0.355
KLM+TRM+Clarity	0.474	0.376	0.424	0.388	0.250	0.341

- As a side note, the agreement among the subjects (for TD)

Kappa	ROUGE-1	ROUGE-2	ROUGE-L
0.544	0.600	0.532	0.527

Objective Evaluations: An Example (2/2)

- Results achieved with **supervised methods**

	Text Documents (TD)			Spoken Documents (SD)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
SVM	0.470	0.364	0.426	0.383	0.245	0.342
Ranking SVM	0.490	0.391	0.447	0.388	0.254	0.344
Perceptron	0.487	0.394	0.439	0.393	0.259	0.352
GCLM	0.482	0.386	0.433	0.380	0.250	0.342
DNN	0.506	0.411	0.466	0.411	0.267	0.370
CNN	0.501	0.404	0.459	0.413	0.271	0.370
CNN+RNN	0.530	0.425	0.485	0.413	0.280	0.372

- **DNN**: Deep Neural Networks
- **CNN**: Convolutional Neural Networks
- **CNN+RNN**: Integration of Convolutional Neural Networks and Recurrent Neural Networks

Subjective Evaluations

- Conduct manual evaluation with respect to factors such as content coverage and linguistic quality
 - Factors that affect linguistic quality could be ***focus, readability, fluency/coherence, referential clarity, ease of understanding, appropriateness***, to name just a few
 - For each factor we may adopt five-level grades: 1-very bad; 2-bad; 3-normal; 4-good; 5-very good to score an automatic summary

Again, to avoid susceptibility to bias produced by personal opinions, the evaluation usually involves several assessors or multiple reference summaries.

Conclusions

- Although various ingenious and sophisticated summarization methods have been developed, most of them are far to be perfect with lots of open questions remained to be solved (still in their infancy? there are still a great number of questions to be solved)
 - E.g., content and linguistic quality of automatic summaries
 - (Leverage or ignore?) the cognitive processes and the knowledge of human beings that go into document understanding
 - Automatic summarization has many possible downstream applications of its own, such as information retrieval, document classification and organization, among others
 - One promising research direction is to harness the power of a wide range of machine learning techniques, such as deep neural networks (DNN) and their variants, word/sentence /document embeddings and curriculum learning, to name just a few
-

Thank You!



Appendix: Risk-Aware Automatic Summarization

– Jointly considering Salience, Relevance and Redundancy

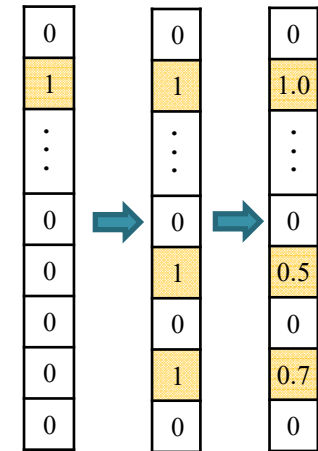
A Risk Minimization Framework (1/4)

- Extractive summarization can be alternatively viewed as a decision making process
 - Select a representative subset of sentences or paragraphs from the original documents → action a
- Bayes decision theory can be employed to guide the summarizer in choosing a course of action
 - It quantifies the tradeoff between
 - Various decisions and the potential cost that accompanies each decision
 - The optimum decision can be made by contemplating each action
 - Choose the action that has the minimum expected risk

1. Lin and Chen, "A risk minimization framework for extractive speech summarization," *ACL 2010*
2. Lin et al., "Extractive speech summarization - From the view of decision theory," *Interspeech 2010*

A Risk Minimization Framework (2/4)

- Without loss of generality, let us denote $\pi \in \Pi$ as a selection strategy
 - It comprises a set of indicators to address the importance of each sentence S_i in a document D to be summarized
 - The feasible selection strategy can be fairly arbitrary according to the underlying principle
 - E.g., **sentence-wise** selection vs. **list-wise** selection (viz. **sentence** selection vs. **summary** selection)



- Moreover, we refer to the k -th action a_k as choosing the k -th selection strategy π_k , and the observation O as the document D

A Risk Minimization Framework (3/4)

- The expected risk of a certain selection strategy π_k

$$R(\pi_k | D) = \int_{\pi} L(\pi_k, \pi) p(\pi | D) d\pi$$

- Therefore, the ultimate goal of extractive summarization could be stated as
 - The search of the best selection strategy π_{opt} from the space of all possible selection strategies that minimizes the expected risk

$$\begin{aligned} \pi_{opt} &= \arg \min_{\pi_k} R(\pi_k | D) \\ &= \arg \min_{\pi_k} \int_{\pi} L(\pi_k, \pi) p(\pi | D) d\pi \end{aligned}$$

A Risk Minimization Framework (4/4)

- Sentence-wise (iterative) selection

$$\begin{aligned}
 S^* &= \arg \min_{S_i \in \tilde{D}} R(S_i | \tilde{D}) \\
 &= \arg \min_{S_i \in \tilde{D}} \sum_{S_j \in \tilde{D}} L(S_i, S_j) P(S_j | \tilde{D})
 \end{aligned}$$

- \tilde{D} denotes the “residual” document
- By applying the Bayes’ rule, the final selection strategy for extractive summarization is stated as

$$S^* = \arg \min_{S_i \in \tilde{D}} \sum_{S_j \in \tilde{D}} L(S_i, S_j) \frac{P(\tilde{D} | S_j) P(S_j)}{\sum_{S_m \in \tilde{D}} P(\tilde{D} | S_m) P(S_m)}$$

Relevance/Redundancy Relevance Significance

 π_2

0
1
⋮
0
0
0
0
0

Relation to Other Summarization Models

- The use of “0-1” loss function

$$S^* = \arg \max_{S_i \in \tilde{D}} \frac{P(\tilde{D} | S_i)P(S_i)}{\sum_{S_m \in \tilde{D}} P(\tilde{D} | S_m)P(S_m)} = \arg \max_{S_i \in \tilde{D}} P(\tilde{D} | S_i)P(S_i)$$

- A natural integration of the supervised and unsupervised summarizers
- Uniform prior distribution
 - Estimate the relevance between the document and sentence using $P(\tilde{D} | S_i)$
- Equal document-likelihood
 - Sentences are selected solely based on the prior probability $P(S_i)$

Implementation Details (1/4)

$$S^* = \arg \min_{S_i \in \tilde{D}} \sum_{S_j \in \tilde{D}} L(S_i, S_j) \frac{P(\tilde{D} | S_j) P(S_j)}{\sum_{S_m \in \tilde{D}} P(\tilde{D} | S_m) P(S_m)}$$

- Sentence Generative Model $P(\tilde{D} | S_i)$
 - We explore the language modeling (LM) approach
 - Each sentence is simply regarded as a probabilistic generative model consisting of a unigram distribution for generating the document

$$P(\tilde{D} | S_i) = \prod_{w \in \tilde{D}} P(w | S_i)^{c(w, \tilde{D})}$$

- Maximum Likelihood Estimation (MLE) of $P(w | S_i)$
 - It may suffer from the problem of unreliable model estimation
 - It can be enhanced with the aid of **topic modeling** (PLSA, LDA, WTM, etc.) or **relevance modeling** or **recurrent neural network modeling**, to name a few

Implementation Details (2/4)

$$S^* = \arg \min_{S_j \in \tilde{D}} \sum_{S_j \in \tilde{D}} L(S_i, S_j) \frac{P(\tilde{D} | S_j) P(S_j)}{\sum_{S_m \in \tilde{D}} P(\tilde{D} | S_m) P(S_m)}$$

- Sentence Prior Model $P(S_i)$
 - We assume the sentence prior probability is in proportion to the posterior probability of a sentence being included in the summary class

$$P(S_i) \approx \frac{p(X_i | \mathbf{S})P(\mathbf{S})}{P(X_i | \mathbf{S})P(\mathbf{S}) + P(X_i | \bar{\mathbf{S}})P(\bar{\mathbf{S}})}$$

- \mathbf{S} and $\bar{\mathbf{S}}$: summary and non-summary classes
- X_i : a set of indicative (prosodic/lexical/structural) features used for representing sentence S_i
- Several popular supervised classifiers can be leveraged for this purpose
 - Bayesian Classifier (BC), Support Vector Machine (SVM), etc.

Implementation Details (3/4)

$$S^* = \arg \min_{S_j \in \tilde{D}} \sum_{S_j \in \tilde{D}} L(S_i, S_j) \frac{P(\tilde{D} | S_j) P(S_j)}{\sum_{S_m \in \tilde{D}} P(\tilde{D} | S_m) P(S_m)}$$

- A rich set of 28 indicative features used to characterize a spoken sentence S_i for modeling $p(X_i | \mathbf{S})$

Structural features	1.Duration of the current sentence 2.Position of the current sentence 3.Length of the current sentence
Lexical Features	1.Number of named entities 2.Number of stop words 3.Bigram language model scores 4.Normalized bigram scores
Acoustic Features	1.The 1st formant 2.The 2nd formant 3.The pitch value 4.The peak normalized cross-correlation of pitch
Relevance Feature	1.VSM score

Implementation Details (4/4)

$$S^* = \arg \min_{S_i \in \tilde{D}} \sum_{S_j \in \tilde{D}} L(S_i, S_j) \frac{P(\tilde{D} | S_j) P(S_j)}{\sum_{S_m \in \tilde{D}} P(\tilde{D} | S_m) P(S_m)}$$

- Loss Function

- VSM-based loss function $L(S_i, S_j)$

- We use the “TF-IDF” weighting to calculate the cosine similarity
- If a sentence is more dissimilar from most of the other sentences, it may incur a higher loss

$$L(S_i, S_j) = 1 - \text{SIM}(S_i, S_j)$$

- MMR-based loss function

- Additionally address the “redundancy” issue

$$L(S_i, S_j) = 1 - \left[\beta \cdot \text{SIM}(S_i, S_j) - (1 - \beta) \cdot \max_{S' \in \text{Summ}} \text{SIM}(S_i, S') \right]$$

- **Summ** the set of already selected summary sentences

Summarization Experiments (1/4)

- MATBN corpus
 - A subset of 205 broadcast news documents was reserved for the summarization experiments
 - 100 documents for training and 20 documents for test
 - The average Chinese character error rate (CER) is about 35%
 - Three subjects were asked to create summaries of the 205 spoken documents
 - The assessment of summarization performance is based on the widely-used ROUGE measure

	ROGUE-1	ROUGE-2	ROUGE-L
Agreement	0.600	0.532	0.527

*The agreement among the subjects for important sentence ranking for the evaluation set.

Summarization Experiments (2/4) $S^* = \arg \min_{S_i \in \tilde{D}} \sum_{S_j \in \tilde{D}} L(S_i, S_j) \frac{P(\tilde{D} | S_j)P(S_j)}{\sum_{S_m \in \tilde{D}} P(\tilde{D} | S_m)P(S_m)}$

- Baseline experiments

- Supervised summarizer – A Bayesian classifier (BC) with 28 indicative features determines the sentence prior probability $P(S_i)$
- Unsupervised summarizer – A (unigram) language modeling approach determines the document-likelihood $P(D | S_i)$

	Text Document (TD)			Spoken Document (SD)		
	ROGUE-1	ROGUE-2	ROGUE-L	ROGUE-1	ROGUE-2	ROGUE-L
BC	0.445 (0.390 - 0.504)	0.346 (0.201 - 0.415)	0.404 (0.348 - 0.468)	0.369 (0.316 - 0.426)	0.241 (0.183 - 0.302)	0.321 (0.268 - 0.378)
LM	0.387 (0.302 - 0.474)	0.264 (0.168 - 0.366)	0.334 (0.251 - 0.415)	0.319 (0.274 - 0.367)	0.164 (0.115 - 0.224)	0.253 (0.215 - 0.301)

- Erroneous transcripts cause significant performance degradation
- BC outperforms LM
 - BC is trained with the handcrafted document-summary data
 - BC utilizes a rich set of features

The results reported here and below are obtained with a setting different from those presented in the aforementioned slides

Summarization Experiments (3/4)

$$S^* = \arg \min_{S_i \in \tilde{D}} \sum_{S_j \in \tilde{D}} L(S_i, S_j) \frac{P(\tilde{D} | S_j) P(S_j)}{\sum_{S_m \in \tilde{D}} P(\tilde{D} | S_m) P(S_m)}$$

- Experiments on proposed methods

		Text Document (TD)			Spoken Document (SD)		
Prior	Loss	ROGUE-1	ROUGE-2	ROUGE-L	ROGUE-1	ROUGE-2	ROUGE-L
	0-1	0.501	0.401	0.459	0.417	0.281	0.356
BC	SIM	0.524	0.425	0.473	0.475	0.351	0.420
	MMR	0.529	0.426	0.479	0.475	0.351	0.420

- Simple “0-1 Loss” gives about 4-5% absolute improvements as compared to the results of BC
- “SIM/MMR Loss” results in higher performance
 - MMR (considering redundancy) is slightly better than SIM
- The performance gaps between the TD and SD cases are reduced to a good extent

Summarization Experiments (4/4) $S^* = \arg \min_{S_i \in \tilde{D}} \sum_{S_j \in \tilde{D}} L(S_i, S_j) \frac{P(\tilde{D} | S_j)P(S_j)}{\sum_{S_m \in \tilde{D}} P(\tilde{D} | S_m)P(S_m)}$

- Experiments on proposed methods

		Text Document (TD)			Spoken Document (SD)		
Prior	Loss	ROGUE-1	ROUGE-2	ROUGE-L	ROGUE-1	ROUGE-2	ROUGE-L
Uniform	SIM	0.405	0.281	0.348	0.365	0.209	0.305
	MMR	0.417	0.282	0.359	0.391	0.236	0.338

- Assume the sentence prior probability $P(S_i)$ is uniformly distributed (**the use of solely unsupervised information**)
 - The importance of a sentence is considered from two angles
 - Relationship between a sentence and the whole document
 - Relationship between the sentence and the other individual sentences
- Additional consideration of the “sentence-sentence” relationship appears to be beneficial

Some Possible Extensions

- Look for different selection strategies
 - E.g., the *listwise* strategy

$$Summary = \arg \min_{\psi_i \in \Psi_D} \sum_{\psi_j \in \Psi_D} L(\psi_i, \psi_j) \frac{P(D | \psi_j) P(\psi_j)}{\sum_{\psi_m \in \Psi_D} P(D | \psi_m) P(\psi_m)}$$

- Explore different modeling approaches and indicative features for the component models
- Investigate discriminative training criteria for training the component models
- Robustly represent the recognition hypotheses of spoken documents beyond the top scoring ones
- Extend and apply the proposed framework to multi-document summarization tasks
- ...

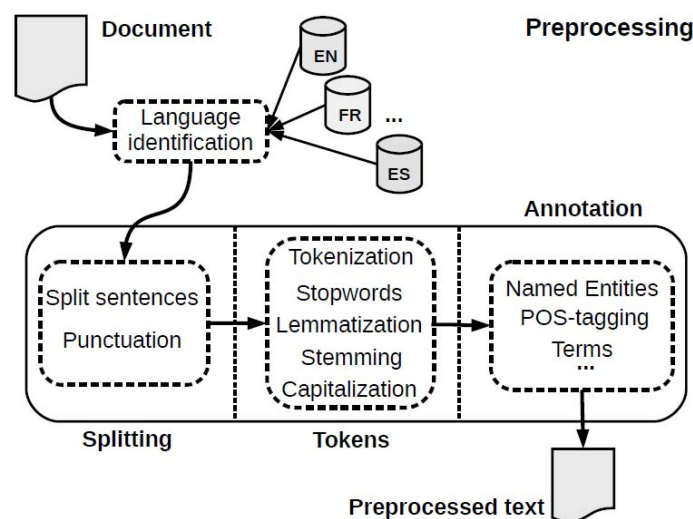
1. Chen and Lin, "A risk-aware modeling framework for speech summarization," IEEE Transactions on Audio, Speech and Language Processing, 2012
2. Chen et al., "Extractive speech summarization using evaluation metric-related training criteria," Information Processing & Management, 2013



Appendix

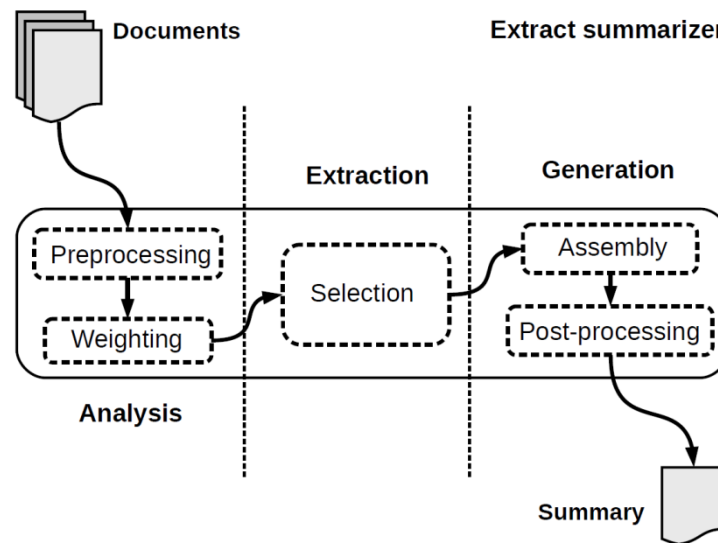
Preprocessing

- *Preprocessing is a difficult task that depends to a large extent on the language in which the text is written*
- **Sentence boundaries**, for example, are demarcated by punctuation, context and quotations, the use of which varies considerably from language to language
- *Moreover, not all languages separate words with spaces. In fact, there are considerable differences between splitting a text written in a European language in Latin characters and an eastern language such as Chinese, Japanese or Korean*



Another Categorization of Extractive Methods

- According to Radev et al. [RAD 02b], algorithms for automatic summarization by extraction can be categorized into three types: surface-level, intermediate-level and deep parsing methods
 - It is a simple idea: a document is split into lexical units (sentences) that are weighted with a heuristic; the units with the highest scores are then extracted and assembled to create a summary



General architecture of an extraction-based summarization system

Surface-level Methods

- Surface-level algorithms do not delve into the linguistic depths of a text; rather they use certain linguistic elements to identify the most relevant segments of a document
 - Used since the very first studies on summarization, surface-level techniques use the occurrences of words to weight sentences [LUH 58]
 - Another surface-level technique is based on the idea that words used in the title are important. Extra weight is, therefore, given to sentences containing these words [EDM 69]
 - Several algorithms use the position of fragments of text. This technique is used for documents with a fixed structure, such as titles, sections, paragraphs and so on. Several studies [BRA 95, LIN 97, HOV 99] have in fact shown that the first lines are always the most important in journalistic articles

Surface-level Methods

- Other algorithms (that mostly use statistical methods but without learning) are based on the properties of the vectorial model for representing texts (VSM) or on graphs
 - For example, CORTEX is a summarization system with a decision algorithm that combines several metrics [TOR 02] (see section 3.7).
 - TEXTRANK and LEXRANK [ERK 04, MIH 04b] represent documents as (directed or undirected) graphs in order to identify relevant sentences (vertices) according to their similarity links

Intermediate-level Methods

- Intermediate-level algorithms use linguistic information that is more sophisticated than surface-level algorithms but less sophisticated than deep parsing
- One intermediate-level technique is lexical chain recognition. Lexical chains are sequences of words connected by lexical semantic relations [BAR 97, SIL 00]
 - In general, producing an automatic summary involves four stages. First, the original text is divided into segments and then lexical chains are constructed. The “strong” chains are identified, and the sentences containing these strong chains are extracted and assembled to create a summary

Deep Parsing Methods

- Deep parsing approaches are based on the idea that it is necessary to use in-depth linguistic techniques that exploit the discursive structure of texts
 - Some of these approaches are based on ***rhetorical structure theory (RST)*** [MAN 87, MAN 88], which aims to finely exploit the structure of the discourse to generate abstracts [ONO 94, MAR 98, MAR 00], or ***on meaning-text theory (MTT)*** [MEL 88, MEL 01]

Sentence Compression and Fusion

- Automatic sentence compression and multisentence fusion (MSF) are two relatively new subjects of research in automatic text summarization
 - Understanding and mastering these topics is important, as they enable a number of improvements to be made, including reducing redundancy and creating summaries that are more similar to human abstracts
- There is a general consensus that sentence fusion and sentence compression are basic forms of paraphrasing
 - They are, therefore, steps toward abstract generation and crucial for improving the quality of this type of summary
- Nevertheless, sentence compression and sentence fusion are entirely different subjects

Sentence Compression

- The idea behind sentence compression is simple: to eliminate all the non-essential information in a sentence while preserving grammaticality
- Turning this idea into an algorithm is, however, a whole other matter, and no method has ever found a definitive solution to this problem
- There is a range of applications for sentence compression:
 - Automatic title generation, generating subtitles for the media, displaying information on mobile devices and so on. Automatic summarization is also concerned with sentence compression. In fact, sentence compression is considered a resource for optimizing automatic summarization

Multisentence Fusion

- The idea of MSF or multisentence compression (MSC) arises from multidocument summarization:
 - The more documents there are about a certain subject, the more important the redundancy in the generated extract.
 - The aim is, therefore, to reduce redundancy while maintaining the essential information contained in a group of semantically similar sentences
- Given that MSF is fundamentally paraphrasing, it can also be used in the context of abstract generation. In fact, one form of basic paraphrasing is the fusion of similar sentences

Cohesion and Coherence

- Extracted sentences are often incoherent and difficult to read. Cohesion and coherence are two ways of directly evaluating synthetically produced summaries, as they are strongly connected to the quality of the summary.
- According to Barzilay and Elhadad [BAR 97]:
 - “Cohesion is a device for ‘sticking together’ different parts of the text. Cohesion is achieved through the use of semantically related terms, co-reference, ellipsis and conjunctions”
- Cohesion is located at the linguistic level of the sentence, whereas coherence is located at the higher level of semantics
 - Cohesion between the sentences of a text is related to the absence of anaphora and unresolved temporal references in the generated summary
 - The coherence of the summary is the absence of contradictions and/or redundancy in the sequence of sentences in a document