A Brief Review of Extractive Summarization Research

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

References:

1. I. Mani and M.T. Maybury (Eds.), *Advances in automatic text summarization*, Cambridge, MA: MIT Press, 1999
2. Document Understanding Conference http://duc.nist.gov/

History of Text Summarization Research

- Research into automatic summarization of text documents dates back to the early1950s
  - However, research work has suffered from a lack of funding for nearly four decades

- Fortunately, the development of the World Wide Web led to a renaissance of the field
  - Summarization was subsequently extended to cover a wider range of tasks, including multi-document, multi-lingual, and multi-media summarization

Spectrum of Text Summarization Research (1/2)

1: Extractive and Abstractive Summarization

- Extractive summarization produces a summary by selecting indicative sentences, passages, or paragraphs from an original document according to a predefined target summarization ratio
- Abstractive summarization provides a fluent and concise abstract of a certain length that reflects the key concepts of the document.
  - This requires highly sophisticated techniques, including semantic representation and inference, as well as natural language generation

In recent years, researchers have tended to focus on extractive summarization.

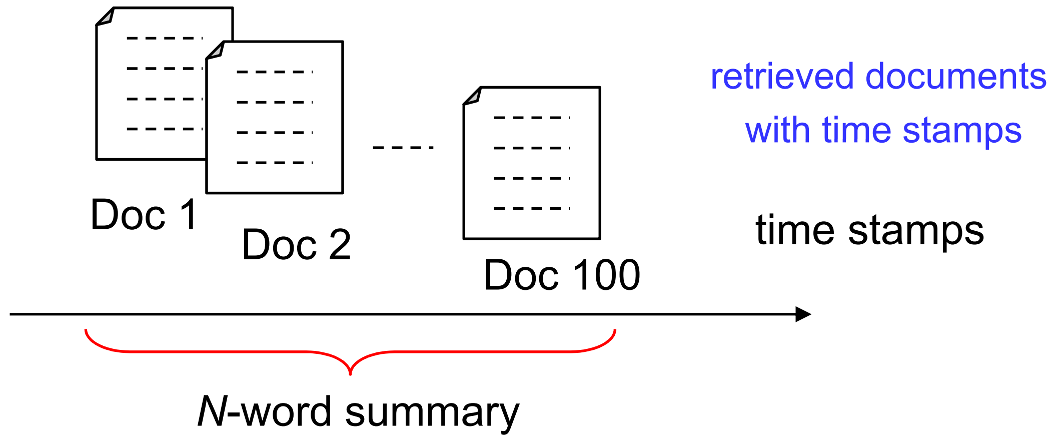
Spectrum of Text Summarization Research (2/2)

2: Generic and Query-oriented Summarization

- A generic summary highlights the most salient information in a document
- A query-oriented summary presents the information in a document that is most relevant to the user's query

Query-oriented (Multi-document) Update Summarization

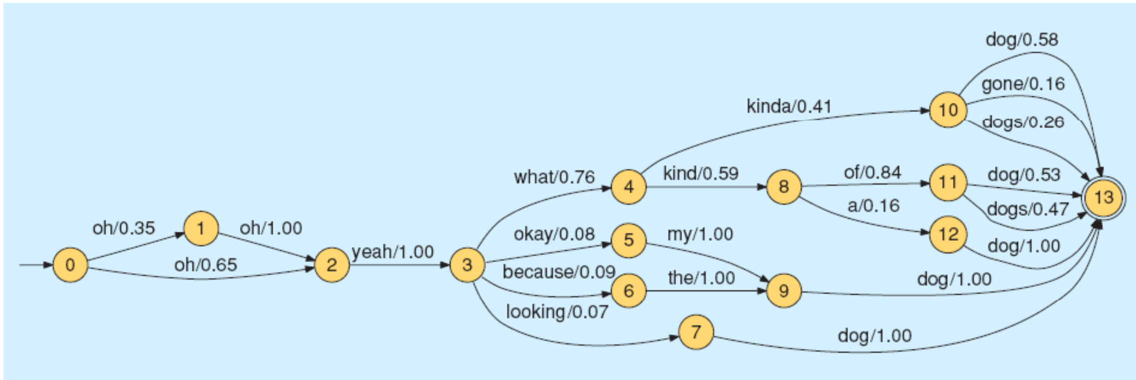
Query: Obama elected president



Special Considerations for Speech Summarization (1/2)

- Speech presents unique difficulties, such as recognition errors, problems with spontaneous speech, and the lack of correct sentence or paragraph boundaries
  - Recognition Errors

word lattice: containing multiple recognition hypotheses



Position-Specific Posterior Probability Lattice (**PSPL**): word position information is readily available

0		1		2		3		4		5		6		7	
Oh	1.0	**Yeah**	.65	**What**	.46	**Kind**	.27	Dog	.26	EOS	.34	**EOS**	.44	EOS	.16
—		Oh	.35	Yeah	.35	What	.27	**Of**	.23	**Dog**	.29	Dog	.09	—	
		—		Because	.06	Kinda	.19	Kind	.16	Dogs	.13	Dogs	.06		
				Okay	.05	The	.06	Kinda	.11	Of	.13	—			
				Looking	.05	My	.05	Dogs	.05	A	.03				
				—		Dog	.05	EOS	.05	Gone	.02				
						......	...	......	...	—					

Cf. Chelba et al. "Retrieval and browsing of spoken content.," *IEEE Signal Processing Magazine* 25 (3), May 2008

Special Considerations for Speech Summarization (2/2)

- Spontaneous effects frequently occur in lectures and conversations
  - Repetitions

    <因為>…<因為> <它> <有><健身><中心>

    because because it has fitness center

  - Hesitations (False starts)

    <台…台灣師範大學>

    Taiwan Normal University

  - Repairs

    <是> <進口> <嗯> <出口> <嗎>

    is import [discourse particle] export [interrogative particle]

  - Filled Pauses

    <我> <去>….<學校>

    I go to school

The first and third examples were adopted from Dr. Che-Kuang Lin's presentation

Typical Features Used for Summarization (1/3)

1. Surface (Structural) Features
   - The position of a sentence in a document or a paragraph
   - The word length in a sentence
   - (For speech) whether an speech utterance is adjacent to a speaker turn

2. Content (Lexical) Features
   - Term frequency (TF) and inversed document frequency (IDF) Scores of the words in a sentence
   - Word *n*-gram (unigram, bigram, etc.) counts of a sentence
   - Number of named entities (such as person names, local names, organization names, dates, artifacts) in a sentence

Typical Features Used for Summarization (2/3)

3. Event Features
   - An event contains event terms and associated event elements
   - Event terms: verbs (such as elect and incorporate) and action nouns (such as election and incorporation) are event terms that can characterize actions
   - Event elements: named entities are considered as event elements, conveying information about “who”, “whom”, “when”, “where”, etc.

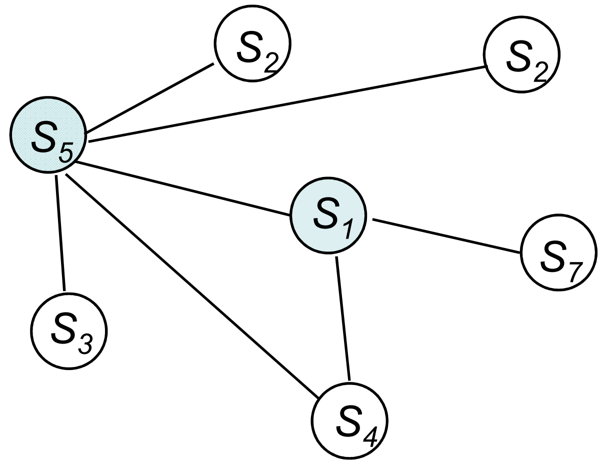
     Barack Hussein Obama was elected the 44th president of the United States on Tuesday

Cf. Wong et al, “Extractive summarization using supervised and unsupervised learning,” *Coling 2008*

Typical Features Used for Summarization (3/3)

4. Relevance Features
   - Sentences highly relevant to the whole document are important
   - Sentences of highly relevant to important sentences are important
   - Sentences related to many other sentences are important (such relationship can be explored by constructing a sentence map or graph and using PageRank (Brin and Page 1998) or HITS (Kleinberg 1999) scores)

HITS: Hyperlink-Induced Topic Search



Graph-based model

5. Acoustic and Prosodic Features (for spoken documents)
   - Energy, pitch, speaking rate
   - Word or sentence duration
   - Recognition confidence score

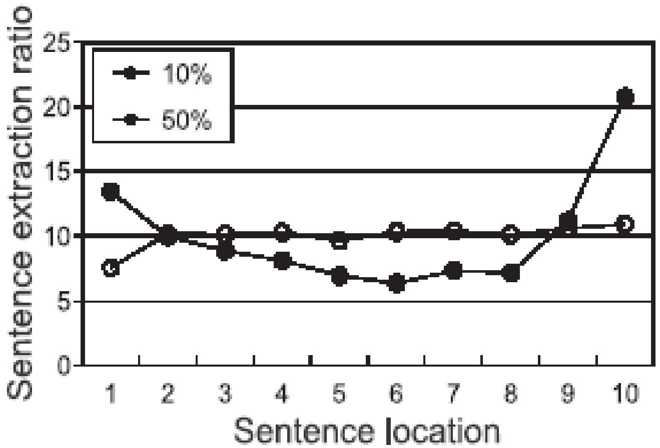
Categorization of Summarization Approaches

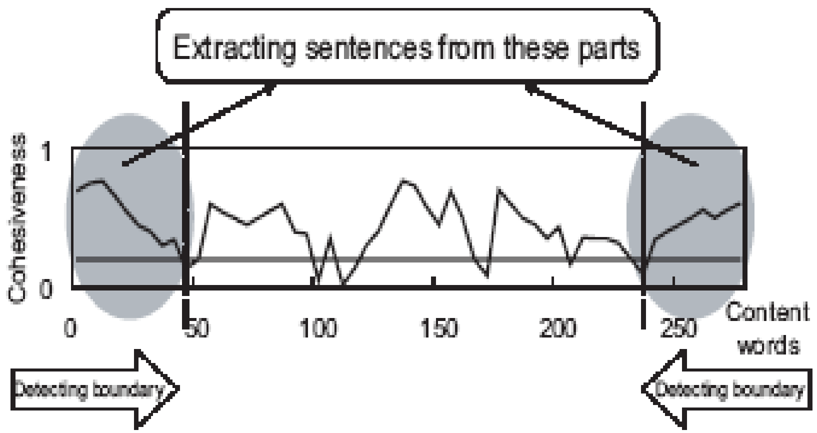
- Unsupervised Summarizers whose models are trained without using handcrafted document-summary pairs
  - Approaches based on sentence structure or location information
  - Approaches based on proximity or significance measures
  - Approaches based on a probabilistic generative framework

- Supervised (Classification-based) Summarizers whose models are trained using handcrafted document-summary pairs
  - Sentence selection is usually formulated as a binary classification problem; that is, a sentence can be included in a summary or omitted
  - Typical models: the Bayesian classifier (BC), the support vector machine (SVM), the conditional random fields (CRF), etc.

Approaches based on Sentence Structure or Location Information

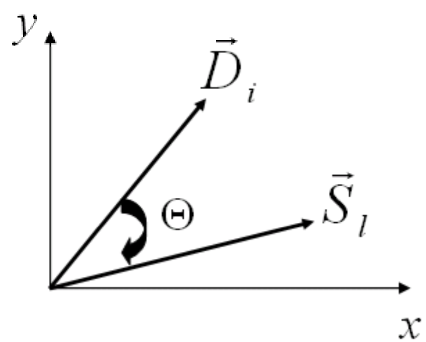
- Lead (Hajime and Manabu 2000) simply chooses the first *N*% of the sentences
- (Hirohata et al. 2005) focuses on the introductory and concluding segments
- (Maskey et al. 2003) selects important sentence based on some specific structures of some domain
  - E.g., broadcast news programs－sentence position, speaker type, previous-speaker type, next-speaker type, speaker change





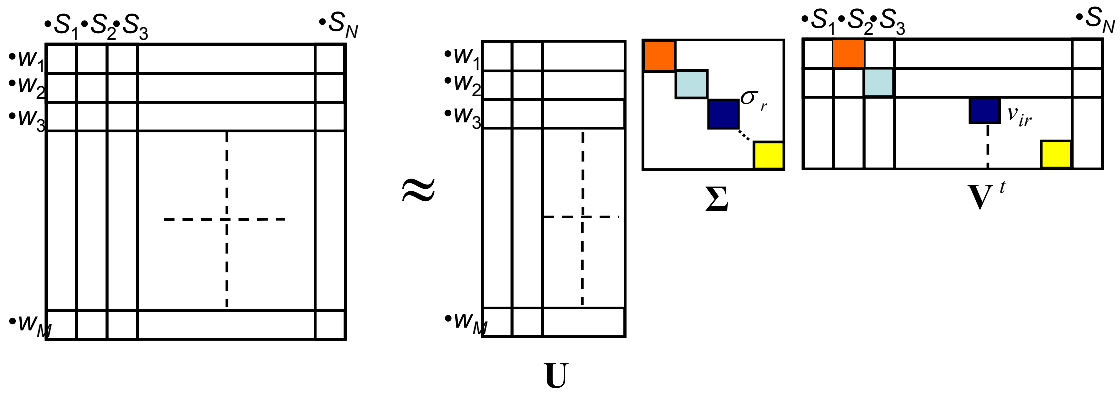
Approaches based on Proximity or Significance Measures (1/4)

- Vector Space Model (VSM) Y. Gong, *SIGIR 2001*
  - Vector representations of sentences and the document to be summarized using statistical weighting such as *TF-IDF*
  - Sentences are ranked based on their proximity to the document
  - To summarize more important and different concepts in a document
    - The terms occurring in the sentence with the highest relevance score $Sim(S_l, D_i)$ are removed from the document
    - The document vector is then reconstructed and the ranking of the rest of the sentences is performed accordingly



Approaches based on Proximity or Significance Measures (2/4)

- Latent Semantic Analysis (LSA) Gong, *SIGIR 2001*
  - 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 semantic vector in the reduced space



  - LSA-1: sentences with the largest index (element) values in each of the top *L* right singular vectors are included in the summary

Approaches based on Proximity or Significance Measures (3/4)

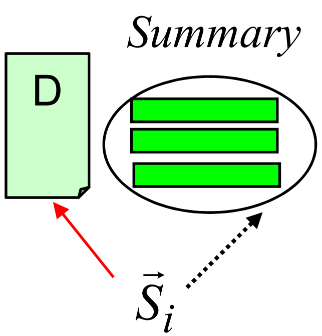
– LSA-2: Sentences also can be selected based on the norms of the semantic vectors (Hirohata et al. 2005)

$$Score(S_i) = \sqrt{\sum_{r=1}^{L} (\sigma_r v_{ir})^2}$$

- Maximal Marginal Relevance (MMR) Carbonell and Goldstien , *S IGIR 1998*
  – Each sentence of a document and the document itself are also represented in vector form, and the cosine score is used for sentence selection
  – Sentence is selected according to two criteria:

    1) whether it is more similar to the whole document than the other sentences, and

    2) whether it is less similar to the set of sentences $\mathbf{S}_l$ selected so far than the other sentences by the following formula



$$NextSen = \max_{S_u} \left[\underbrace{\beta \cdot sim(S_u, D)}_{\text{relevance component}} - \underbrace{(1-\beta) \max_{S_j \in \mathbf{S}_l} sim(S_u, S_j)}_{\text{redundancy component}} \right],$$

Approaches based on Proximity or Significance Measures (4/4)

- Sentence Significance Score (SIG)
  - Sentences are ranked based on their significance which, for example, is defined by the average importance scores of words in the sentence

$$SIG(S_i) = \frac{1}{N_S} \sum_{n=1}^{N_S} I(w_n)$$

$$I(w_n) = f_w \cdot icf = f_w \cdot \log \frac{F_C}{F_w}$$

similar to *TF-IDF* weighting

Furui et al., IEEE SAP 12(4), 2004

  - Other features such as *word confidence*, *linguistic score*, or *prosodic information* also can be further integrated into this method

$$SIG(S_i) = \frac{1}{N_{S_i}} \sum_{n=1}^{N_{S_i}} \{\lambda_1 s(w_n) + \lambda_2 l(w_n) + \lambda_3 c(w_n) + \lambda_4 g(w_n)\} + \lambda_5 b(S_i)$$

    - $s(w_n)$:statistical measure, such as TF/IDF
    - $l(w_n)$:linguistic measure, e.g., named entities and POSs
    - $c(w_n)$:confidence score
    - $g(w_n)$:N-gram score
    - $b(S_i)$:calculated from the grammatical structure of the sentence

Approaches based on a Probabilistic Generative Framework (1/2)

- Criterion: Maximum a posteriori (MAP)

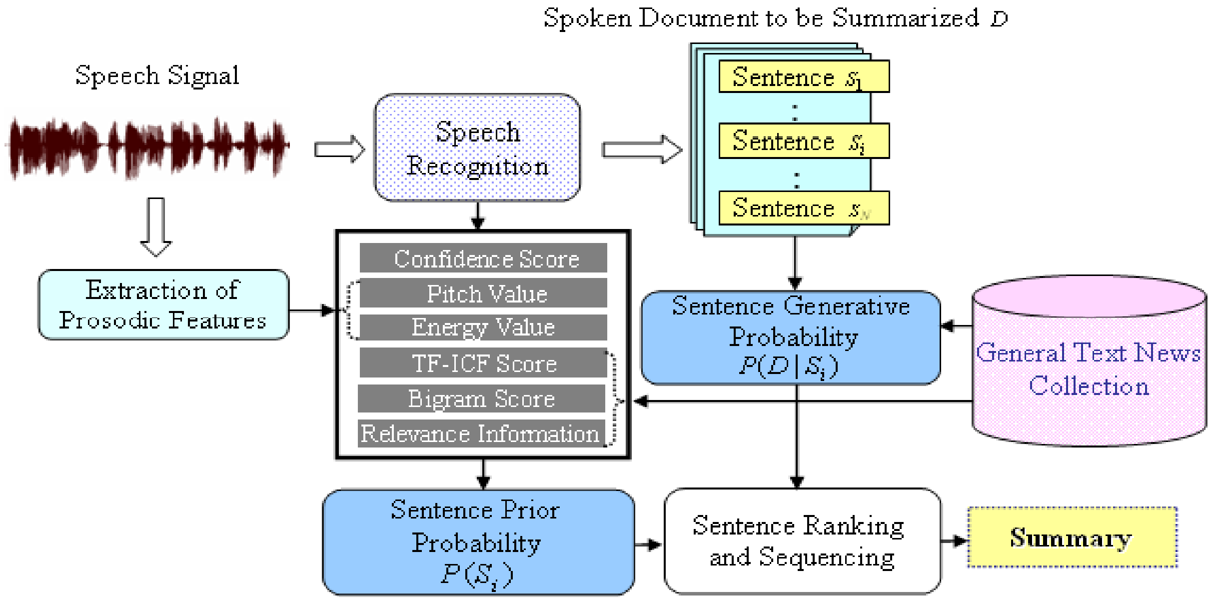
$$P(S_i|D) = \frac{P(D|S_i)P(S_i)}{P(D)} \overset{\text{rank}}{=} P(D|S_i)P(S_i)$$

- Sentence Generative Model, $P(D|S_i)$
  - Each sentence of the document as a probabilistic generative model
  - Language Model (LM), Sentence Topic Model (STM) and Word Topic Model (WTM) are initially investigated
- Sentence Prior Distribution, $P(S_i)$
  - The sentence prior distribution may have to do with sentence duration/position, correctness of sentence boundary, confidence score, prosodic information, etc. (e.g., they can be fused by the whole-sentence maximum entropy model)

Cf. Chen et al., "A probabilistic generative framework for extractive broadcast news speech summarization," to appear in *IEEE Transactions on Audio, Speech and Language Processing*

Approaches based on a Probabilistic Generative Framework (2/2)

- A probabilistic generative framework for speech summarization



  - E.g., the sentence generative model is implemented with the language model (LM) or sentence topic model (STM)

$$P_{\mathrm{LM}}(D|S_i) = \prod_{w_n \in D} \left[\lambda \cdot P(w_n|S_i) + (1-\lambda) \cdot P(w_n|C) \right]^{c(w_n, D)}$$

$$P_{\mathrm{STM}}(D|S_i) = \prod_{w_n \in D} \left[\sum_{k=1}^{K} P(w_n|T_k) P(T_k|S_i) \right]^{c(w_n, D)}$$

Classification-based Summarizers (1/3)

- Extractive document summarization can be treated as a two-class (summary/non-summary) classification problem of a given sentence
  - A sentence with a set of representative features $X_i = \{x_{i1}, \cdots, x_{ij}, \cdots, x_{iJ}\}$ is input to the classifier
  - The important sentences of a document D can be selected (or ranked) based on $P(S_i \in \mathbf{S} \mid X_i)$, the posterior probability of a sentence S_i being included in the summary $\mathbf{S}$ given the feature set X_i
- Bayesian Classifier (BC)

$$P(S_i \in \mathbf{S} \mid X_i) = \frac{p(X_i \mid S_i \in \mathbf{S})P(S_i \in \mathbf{S})}{P(X_i)} \propto p(X_i \mid S_i \in \mathbf{S})P(S_i \in \mathbf{S})$$

  - Naïve Bayesian Classifier (NBC)

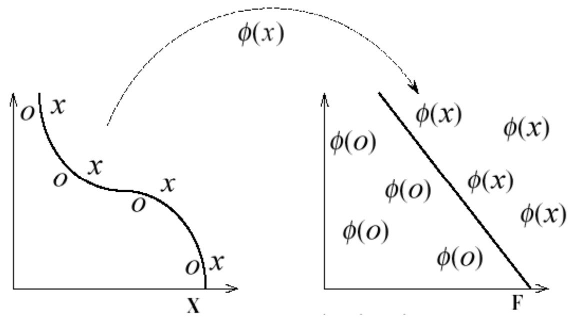
features x_{ij} are conditionally independent given $S_i \in \mathbf{S}$

$$P(S_i \in \mathbf{S} \mid X_i) = P(S_i \in \mathbf{S} \mid x_{i1}, \ldots, x_{ij}, \ldots, x_{iJ}) \propto P(S_i \in \mathbf{S}) \prod_{j=1}^{J} P(x_{ij} \mid S_i \in \mathbf{S})$$

Classification-based Summarizers (2/3)

- Support Vector Machine (SVM)
  – SVM is expected to find a hyper-plane to separate sentences of the document as summary or non-summary sentence

$$y_i\left(w^T\phi\left(X_i\right)+b\right)\geq 1-\xi_i$$



$$P\left(S_i\in\mathbf{S}\middle|X_i\right)\approx\frac{1}{1+\exp\left(\alpha\cdot\left(w^T\phi\left(X_i\right)+b\right)+\beta\right)}$$

Classification-based Summarizers (3/3)

- Conditional Random Fields
  - CRF can effectively capture the dependent relationships among sentences
    - CRF is an undirected discriminative graphical model that combines the advantages of the maximum entropy Markov model (MEMM) and the hidden Markov model (HMM)

$$p(\mathbf{Y} \mid \mathbf{X}) = \frac{1}{Z_{\mathbf{X}}} \exp\left(\sum_{i=1}^{I} \sum_{k} \lambda_k f_k(y_i, X_i) \right)$$

$\mathbf{X} = \{X_1, \ldots, X_i, \ldots, X_I\}$: the entire sentence sequence of a document

$\mathbf{Y} = \{y_1, \ldots, y_i, \ldots, y_I\}$: state sequence, where each y_i can be a summary or non-summary state

$f_k(y_i, X_i)$: a function that measures a feature relating the state y_i for sentence S_i with the input features X_i

λ_i : the weight of each feature function

Evaluation Metrics (1/2)

- Subjective Evaluation Metrics (direct evaluation)
  - Conducted by human subjects
  - Different levels
- Objective Evaluation Metrics
  - Automatic summaries were evaluated by objective metrics
- Automatic Evaluation
  - Summaries are evaluated by IR

Evaluation Metrics (2/2)

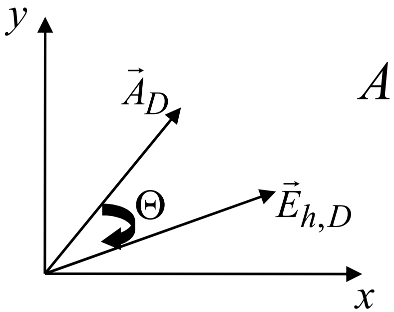
- Objective Evaluation Metrics
  - **ROUGE-*N*** (Lin et al. 2003)
    - ROUGE-*N* is an *N*-gram recall between an automatic summary and a set of manual summaries

$$\text{ROUGE}-N = \frac{\sum_{\mathbf{S}\in\mathbf{S}_H}\sum_{g_N\in\mathbf{S}} C_m(g_N)}{\sum_{\mathbf{S}\in\mathbf{S}_H}\sum_{g_N\in\mathbf{S}} C(g_N)}$$

S_H : a set of human summaries

$C_m(g_N)$: number of matched N-grams between human and automatic summary

  - **Cosine Measure** (Saggion et al. 2002)



$$Acc_D = \frac{1}{2}\left[sim\left(E, E_R\right) + sim\left(E, A_R\right)\right]$$

E : automatic extractive summary

E_R : reference extractive summary

A_R : reference abstractiv e summary

Experimental Results (1/4)

- Preliminary tests on 205 broadcast news stories (100: development; 105:) collected in Taiwan (automatic transcripts with 30% character error rate)
  - ROUGE-2 scores for supervised summarizers

		Summarization Ratio		
		10%	20%	30%
BC	TD	0.490	0.583	0.589
	SD	0.321	0.331	0.317
SVM	TD	0.545	0.625	0.637
	SD	0.333	0.363	0.353
CRF	TD	0.547	0.654	0.637
	SD	0.346	0.371	0.364

TD: manual transcription of broadcast news documents

SD: automatic transcription of broadcast news documents by speech recognition

Cf. lin et al., "A comparative study of probabilistic ranking models for Chinese spoken document summarization," to appear in *ACM Transactions on Asian Language Information Processing,* March 2009

Experimental Results (2/4)

– ROUGE-2 scores for unsupervised summarizers

		Summarization Ratio		
		10%	20%	30%
VSM	TD	0.286	0.427	0.492
	SD	0.204	0.239	0.282
LSA	TD	0.213	0.325	0.418
	SD	0.187	0.240	0.276
MMR	TD	0.292	0.433	0.492
	SD	0.204	0.241	0.280
SIG	TD	0.248	0.408	0.450
	SD	0.179	0.213	0.248
LM	TD	0.328	0.450	0.501
	SD	0.201	0.250	0.282
STM	TD	0.335	0.453	0.494
	SD	0.211	0.262	0.286
RND	TD	0.110	0.188	0.289
	SD	0.163	0.223	0.230

Experimental Results (3/4)

– ROUGE-2 scores for supervised summarizers trained without manual labeling (i.e., STM Labeling +Data Selection and STM Labeling)

	STM Labeling + Data Selection		STM Labeling		Manual Labeling	
	SVM	CRF	SVM	CRF	SVM	CRF
10%	0.232	0.283	0.165	0.194	0.333	0.346
20%	0.262	0.275	0.253	0.262	0.363	0.371
30%	0.291	0.295	0.291	0.296	0.353	0.364

- Data selection using sentence relevance information

$$avgSim\left(S_i\right)=\frac{\sum_{D_l \in \mathbf{D}^i_{\mathrm{top}\,M}} \sum_{\substack{D_u \in \mathbf{D}^i_{\mathrm{top}\,M} \\ D_l \neq D_u}} \frac{\vec{D}_l \cdot \vec{D}_u}{\left\|\vec{D}_l\right\| \cdot \left\|\vec{D}_u\right\|}}{M \cdot \left(M-1\right)}$$

	10%	20%	30%
Summary sentences	0.059	0.057	0.055
Non-summary sentences	0.047	0.046	0.045

Experimental Results (4/4)

- Analysis of features' contributions to summarization performance (CRF taken as an example)

		Summarization Ratio		
		10%	20%	30%
Ac	TD	0.425	0.567	0.574
	SD	0.315	0.336	0.321
St	TD	0.369	0.458	0.490
	SD	0.144	0.132	0.159
Le	TD	0.324	0.464	0.494
	SD	0.287	0.272	0.273
Re	TD	0.391	0.486	0.529
	SD	0.284	0.302	0.313
Ac + St	TD	0.501	0.609	0.621
	SD	0.327	0.350	0.345
Le + Re	TD	0.510	0.555	0.577
	SD	0.302	0.318	0.319
Ac + St + Le	TD	0.495	0.634	0.622
	SD	0.319	0.368	0.343
Ac + St + Re	TD	0.545	0.631	0.634
	SD	0.346	0.362	0.350
Ac + St + Le + Re	TD	0.547	0.654	0.637
	SD	0.346	0.371	0.364
Ac + St + Le + Re + Ge	TD	0.595	0.657	0.644
	SD	0.351	0.372	0.369

Detailed Information of the Features Used for Summarization

St	Structural features	*POSITION*: Sentence position *DURATION*: Duration of the preceding/current/following sentence
Le	Lexical Features	*BIGRAM_SCORE*: Normalized bigram language model scores *SIMILARITY*: Similarity scores between a sentence and its preceding/following neighbor sentence *NUM_NAME_ENTITIES*: Number of named entities (NEs) in a sentence
Ac	Acoustic Features	*PITCH*: Min/max/mean/difference pitch values of a spoken sentence *ENERGY*: Min/max/mean/difference value of energy features of a spoken sentence *CONFIDENCE*: Posterior probabilities
Re	Relevance Features	*R-VSM*: Relevance score obtained by using the VSM summarizer *R-LSA*: Relevance score obtained by using the LSA summarizer

Ge: the scores derived by LM and STM