Automatic Summarization

Ani Nenkova University of Pennsylvania
Sameer Maskey IBM Research
Yang Liu University of Texas at Dallas

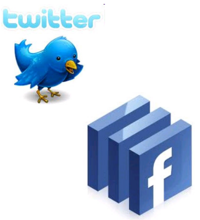
Why summarize?

Text summarization

News articles

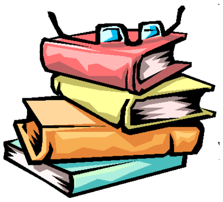
© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com
"I'll explain it one more time. Each of these beeps is a heartbeat, not an incoming e-mail."

Emails



Social Media Streams

Scientific Articles



Books

Websites

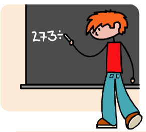
Speech summarization

Phone Conversation

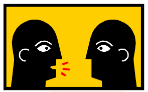
Lecture

Meeting

Talk Shows



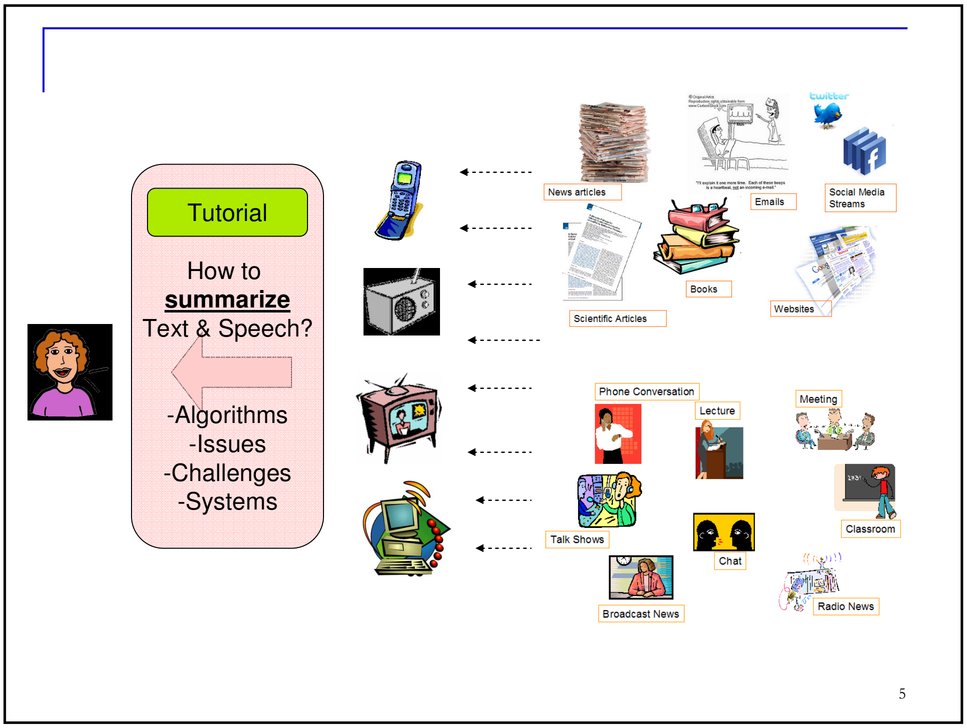
Classroom



Chat

Broadcast News

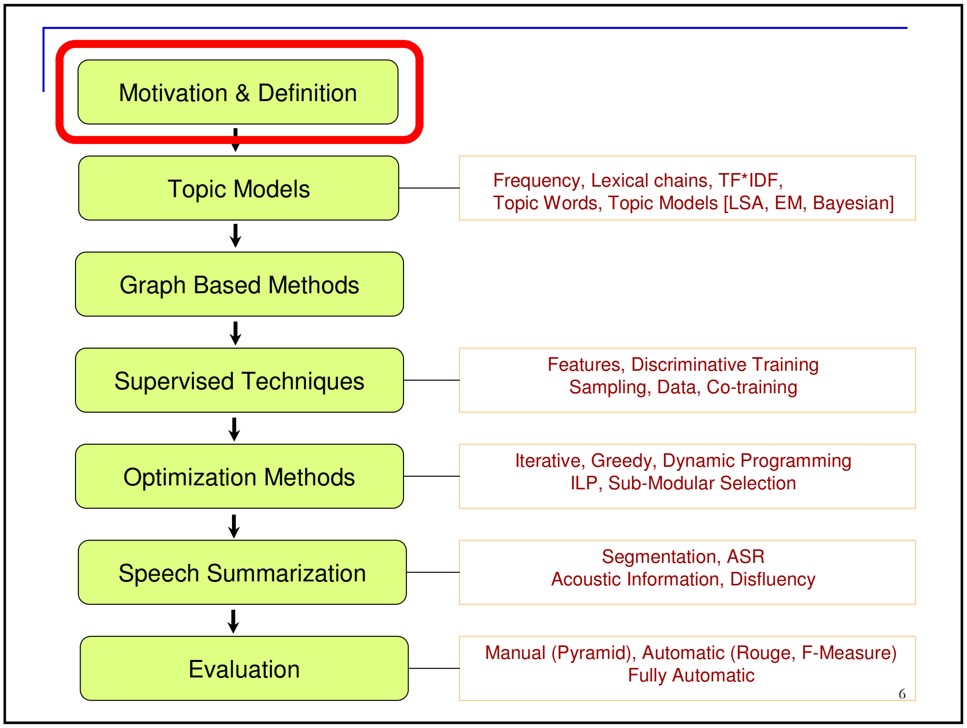
Radio News



Tutorial

How to **summarize** Text & Speech?

-Algorithms
-Issues
-Challenges
-Systems



Motivation & Definition

Topic Models — Frequency, Lexical chains, TF*IDF, Topic Words, Topic Models [LSA, EM, Bayesian]

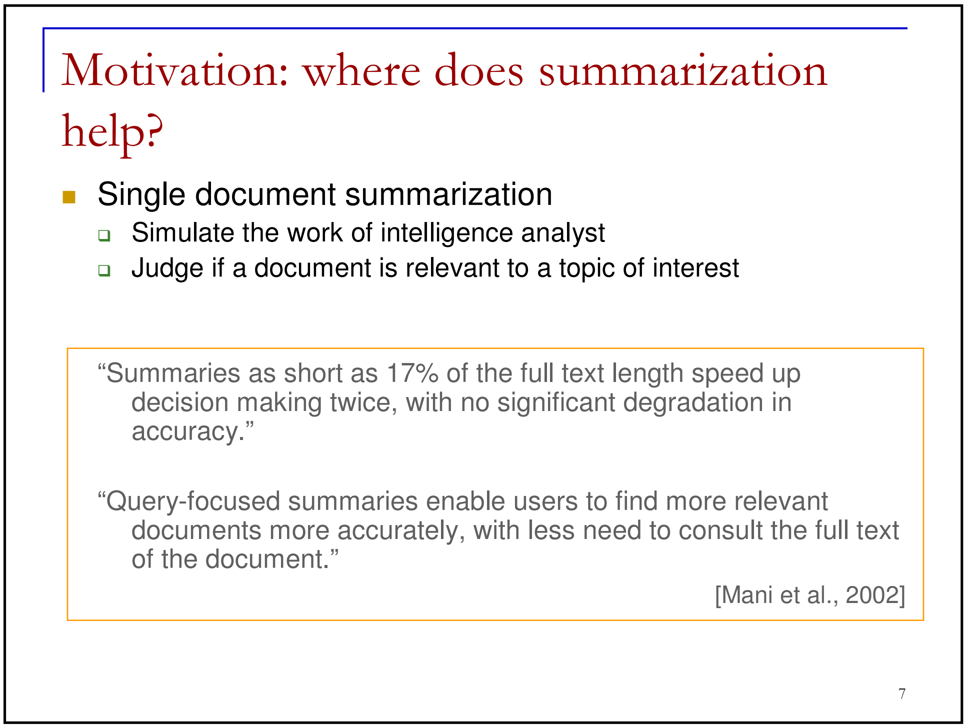
Graph Based Methods

Supervised Techniques — Features, Discriminative Training Sampling, Data, Co-training

Optimization Methods — Iterative, Greedy, Dynamic Programming ILP, Sub-Modular Selection

Speech Summarization — Segmentation, ASR Acoustic Information, Disfluency

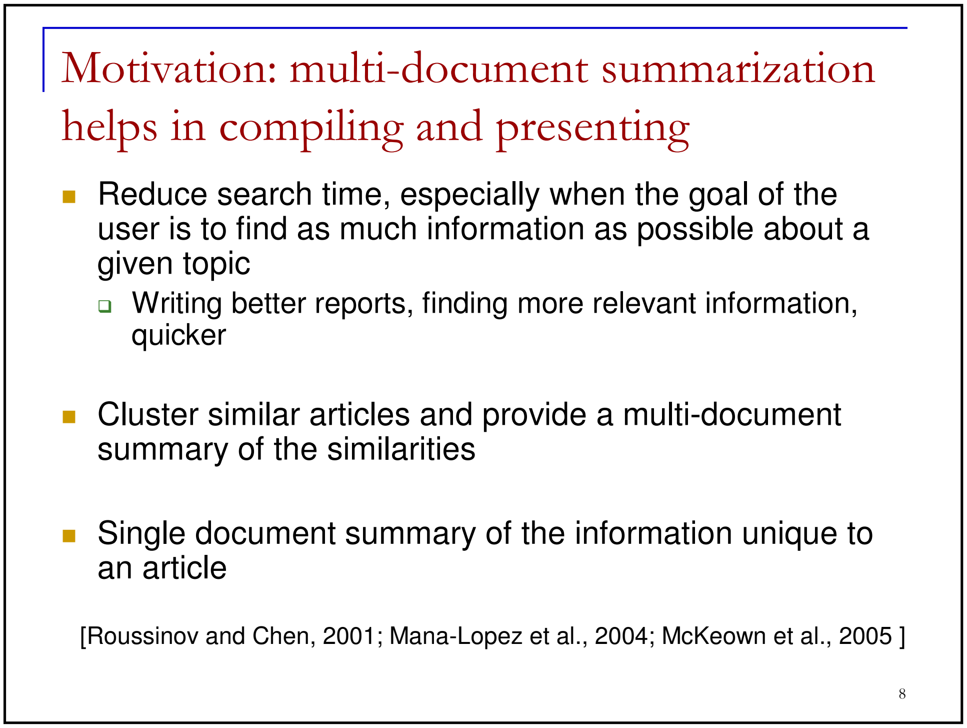
Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure) Fully Automatic



Motivation: where does summarization help?

- Single document summarization
  - Simulate the work of intelligence analyst
  - Judge if a document is relevant to a topic of interest

> "Summaries as short as 17% of the full text length speed up decision making twice, with no significant degradation in accuracy."
>
> "Query-focused summaries enable users to find more relevant documents more accurately, with less need to consult the full text of the document."
>
> [Mani et al., 2002]



Motivation: multi-document summarization helps in compiling and presenting

- Reduce search time, especially when the goal of the user is to find as much information as possible about a given topic
  - Writing better reports, finding more relevant information, quicker
- Cluster similar articles and provide a multi-document summary of the similarities
- Single document summary of the information unique to an article

[Roussinov and Chen, 2001; Mana-Lopez et al., 2004; McKeown et al., 2005]

Benefits from speech summarization

- Voicemail
  - Shorter time spent on listening (call centers)
- Meetings
  - Easier to find main points
- Broadcast News
  - Summary of story from mulitiple channels
- Lectures
  - Useful for reviewing of course materials

[He et al., 2000; Tucker and Whittaker, 2008; Murray et al., 2009]

Assessing summary quality: overview

- Responsiveness
  - Assessor directly rate each summary on a scale
  - In official evaluations but rarely reported in papers
- Pyramid
  - Assessors create model summaries
  - Assessors identifies semantic overlap between summary and models
- ROUGE
  - Assessors create model summaries
  - ROUGE automatically computes word overlap

Tasks in summarization

Content (sentence) selection

- Extractive summarization

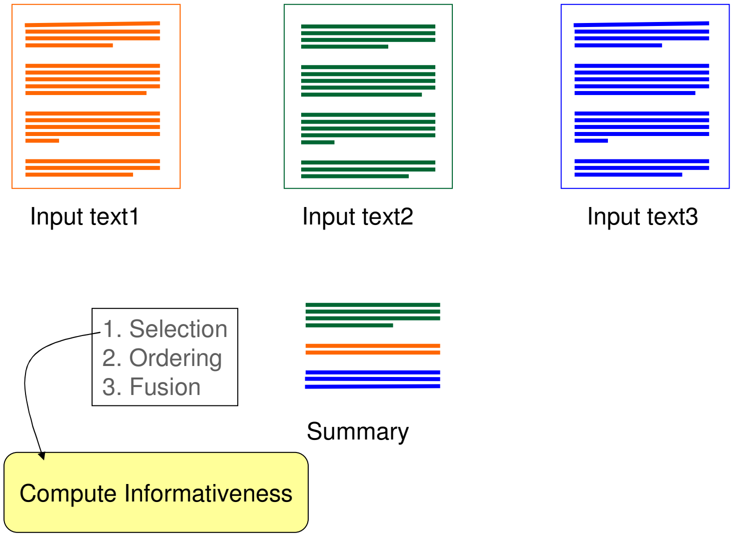
Information ordering

- In what order to present the selected sentences, especially in multi-document summarization

Automatic editing, information fusion and compression

- Abstractive summaries

Extractive (multi-document) summarization



Input text1

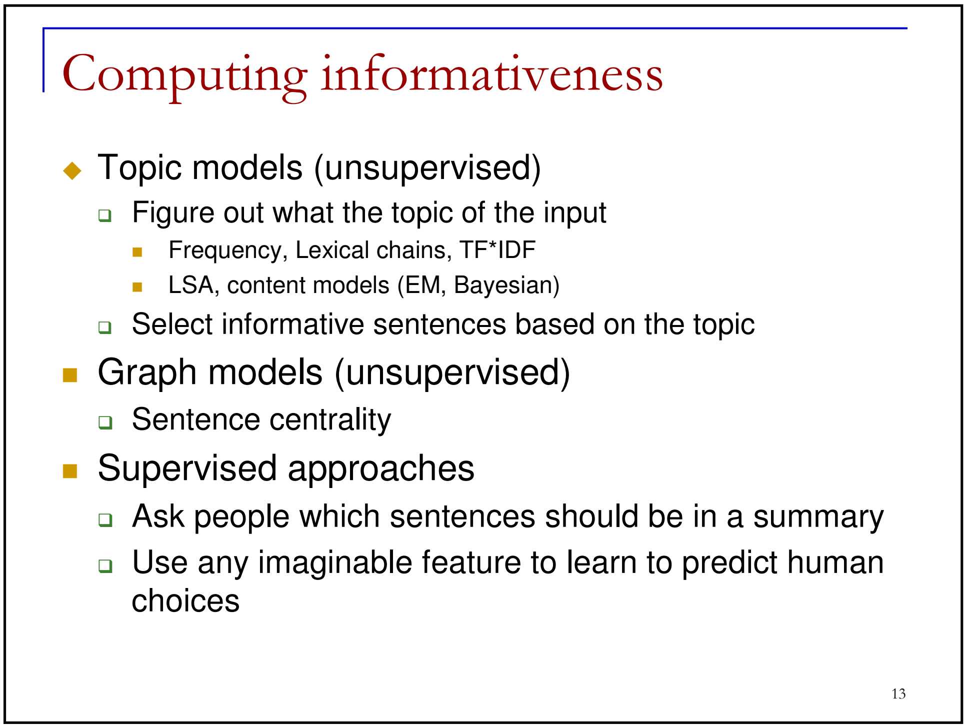
Input text2

Input text3

1. Selection
2. Ordering
3. Fusion

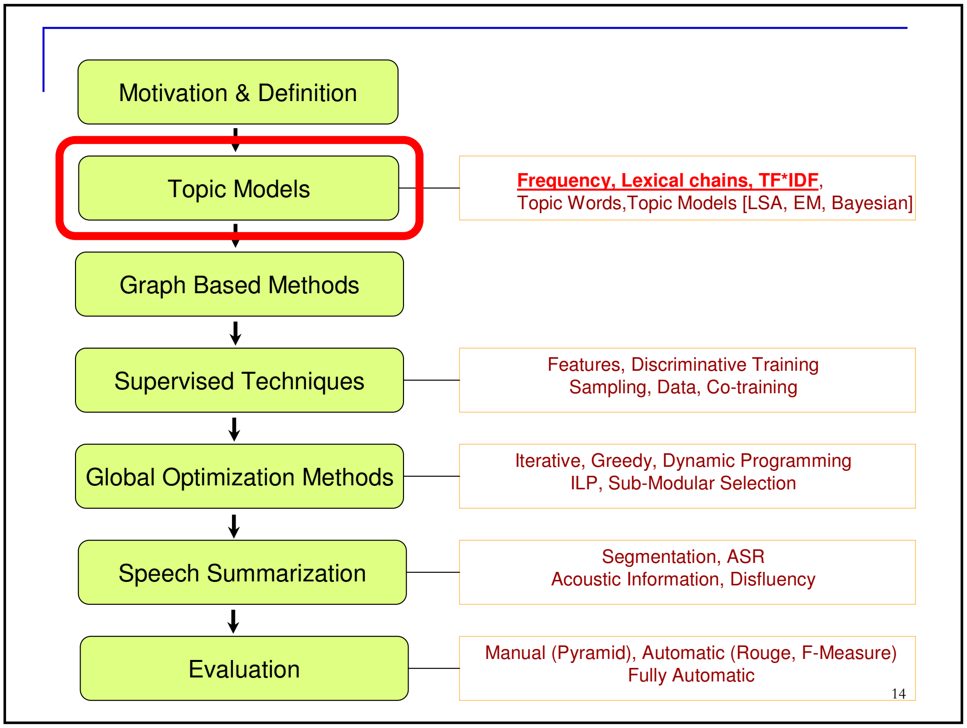
Summary

Compute Informativeness



Computing informativeness

- Topic models (unsupervised)
  - Figure out what the topic of the input
    - Frequency, Lexical chains, TF*IDF
    - LSA, content models (EM, Bayesian)
  - Select informative sentences based on the topic
- Graph models (unsupervised)
  - Sentence centrality
- Supervised approaches
  - Ask people which sentences should be in a summary
  - Use any imaginable feature to learn to predict human choices



Motivation & Definition

Topic Models — **Frequency, Lexical chains, TF*IDF**, Topic Words,Topic Models [LSA, EM, Bayesian]

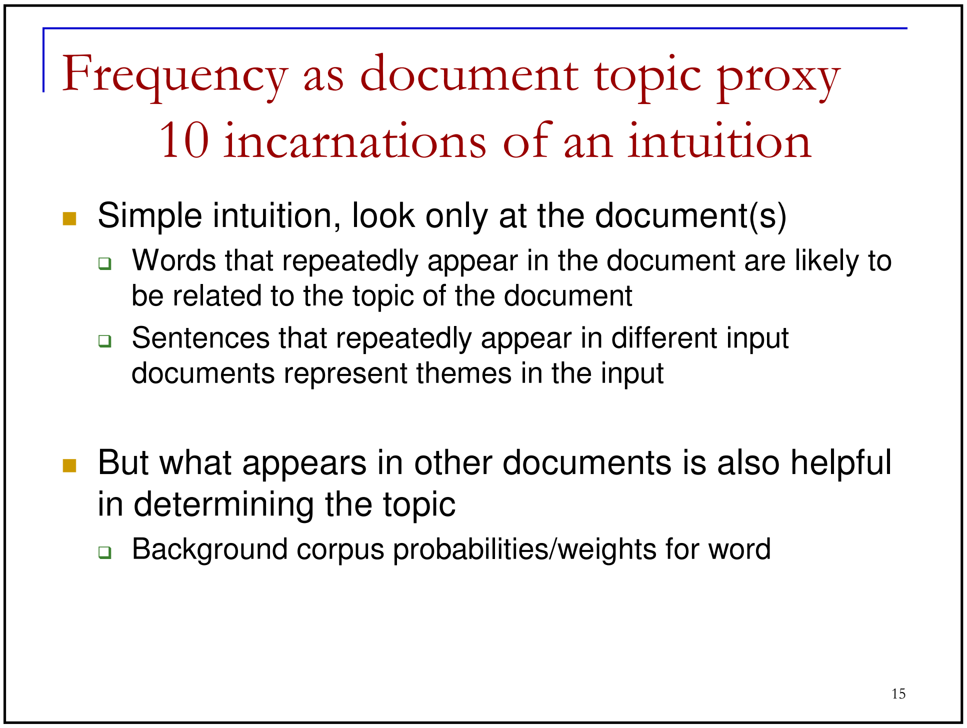
Graph Based Methods

Supervised Techniques — Features, Discriminative Training Sampling, Data, Co-training

Global Optimization Methods — Iterative, Greedy, Dynamic Programming ILP, Sub-Modular Selection

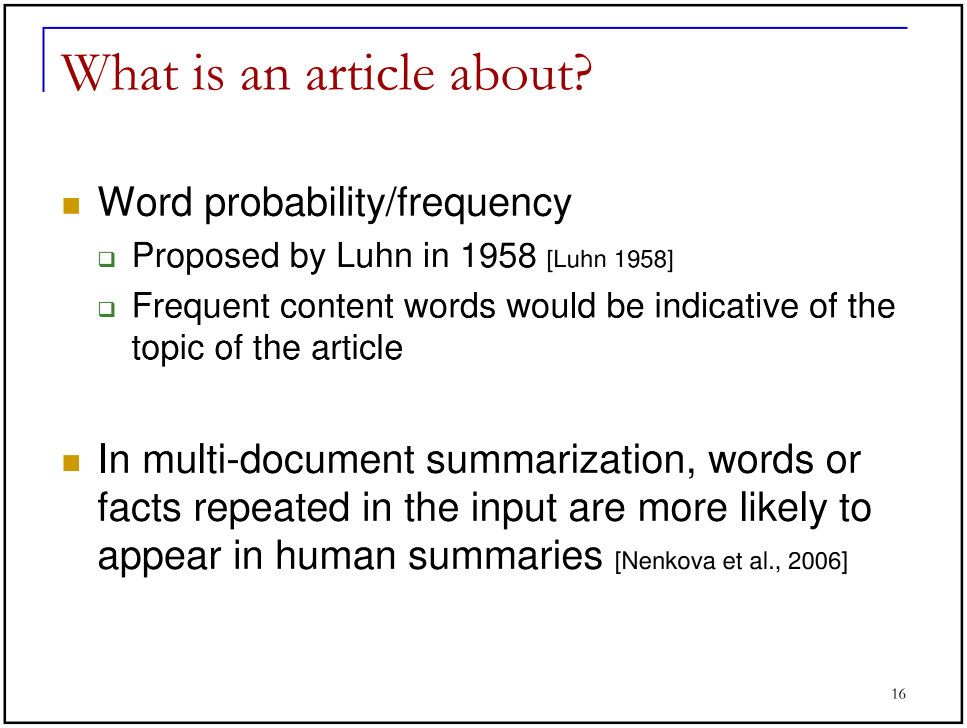
Speech Summarization — Segmentation, ASR Acoustic Information, Disfluency

Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure) Fully Automatic



Frequency as document topic proxy 10 incarnations of an intuition

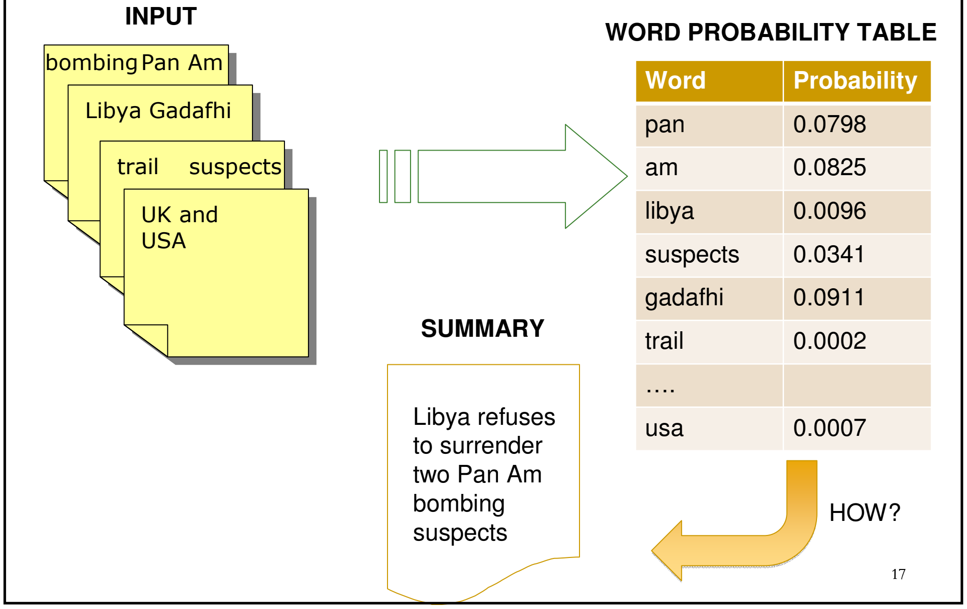
- Simple intuition, look only at the document(s)
  - Words that repeatedly appear in the document are likely to be related to the topic of the document
  - Sentences that repeatedly appear in different input documents represent themes in the input
- But what appears in other documents is also helpful in determining the topic
  - Background corpus probabilities/weights for word



What is an article about?

- Word probability/frequency
  - Proposed by Luhn in 1958 [Luhn 1958]
  - Frequent content words would be indicative of the topic of the article
- In multi-document summarization, words or facts repeated in the input are more likely to appear in human summaries [Nenkova et al., 2006]

Word probability/weights



WORD PROBABILITY TABLE

Word	Probability
pan	0.0798
am	0.0825
libya	0.0096
suspects	0.0341
gadafhi	0.0911
trail	0.0002
….	
usa	0.0007

HOW: Main steps in sentence selection according to word probabilities

Step 1 Estimate word weights (probabilities)

Step 2 Estimate sentence weights

$$Weight(Sent) = CF(w_i \in Sent)$$

Step 3 Choose best sentence

Step 4 Update word weights

Step 5 Go to 2 if desired length not reached

More specific choices [Vanderwende et al., 2007; Yih et al., 2007; Haghighi and Vanderwende, 2009]

- Select highest scoring sentence

$$Score(S) = \frac{1}{|S|}\sum_{w \in S} p(w)$$

- Update word probabilities for the selected sentence to reduce redundancy

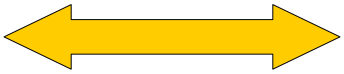
$$p^{new}(w) = p^{old}(w).p^{old}(w)$$

- Repeat until desired summary length

Is this a reasonable approach: yes, people seem to be doing something similar

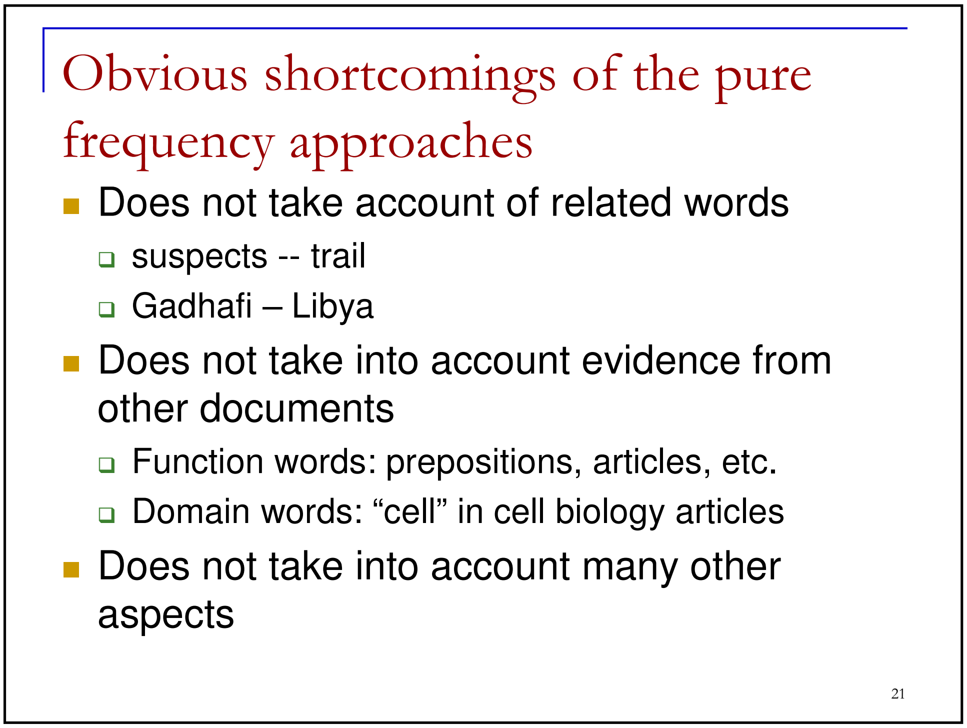
- Simple test
  - Compute word probability table from the input
  - Get a batch of summaries written by H(umans) and S(ystems)
  - Compute the likelihood of the summaries given the word probability table
- Results
  - Human summaries have higher likelihood

LOW



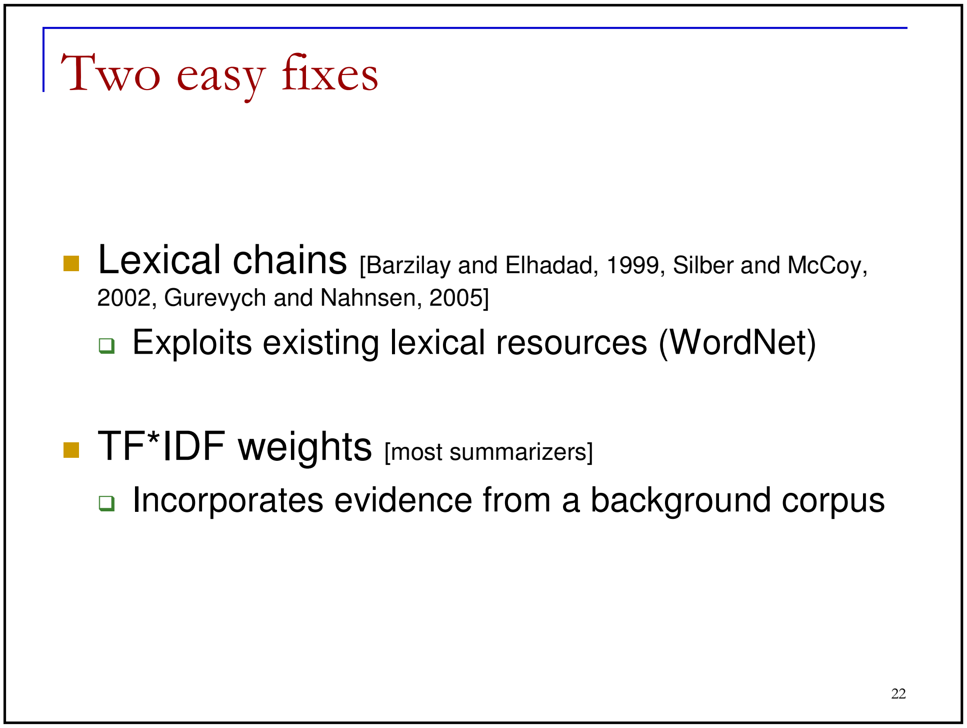
HIGH LIKELIHOOD

HSSSSSSSSSSHSSSHSSHHSHHHHH



Obvious shortcomings of the pure frequency approaches

- Does not take account of related words
  - suspects -- trail
  - Gadhafi – Libya
- Does not take into account evidence from other documents
  - Function words: prepositions, articles, etc.
  - Domain words: “cell” in cell biology articles
- Does not take into account many other aspects



Two easy fixes

- Lexical chains [Barzilay and Elhadad, 1999, Silber and McCoy, 2002, Gurevych and Nahnsen, 2005]
  - Exploits existing lexical resources (WordNet)
- TF*IDF weights [most summarizers]
  - Incorporates evidence from a background corpus

Lexical chains and WordNet relations

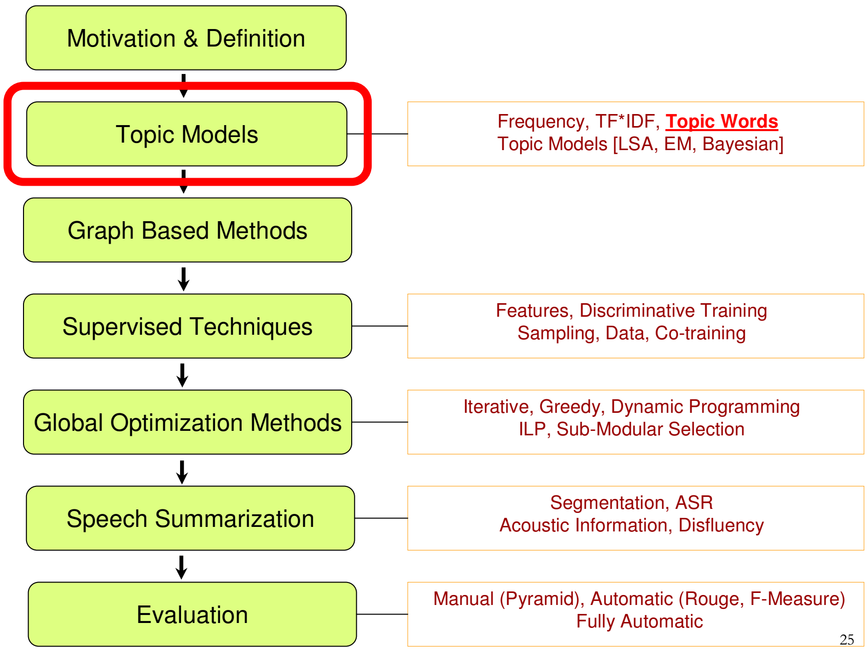
- Lexical chains
  - Word sense disambiguation is performed
  - Then topically related words represent a topic
    - Synonyms, hyponyms, hypernyms
  - Importance is determined by frequency of the words in a topic rather than a single word
  - One sentence per topic is selected
- Concepts based on WordNet [Schiffman et al., 2002, Ye et al., 2007]
  - No word sense disambiguation is performed
    - {war, campaign, warfare, effort, cause, operation}
    - {concern, carrier, worry, fear, scare}

TF*IDF weights for words

Combining evidence for document topics from the input and from a background corpus

- Term Frequency (TF)
  - Times a word occurs in the input
- Inverse Document Frequency (IDF)
  - Number of documents (df) from a background corpus of *N* documents that contain the word

$$\boxed{TF * IDF = tf \times \log(N / df)}$$



Motivation & Definition

Topic Models — Frequency, TF*IDF, **Topic Words**
Topic Models [LSA, EM, Bayesian]

Graph Based Methods

Supervised Techniques — Features, Discriminative Training
Sampling, Data, Co-training

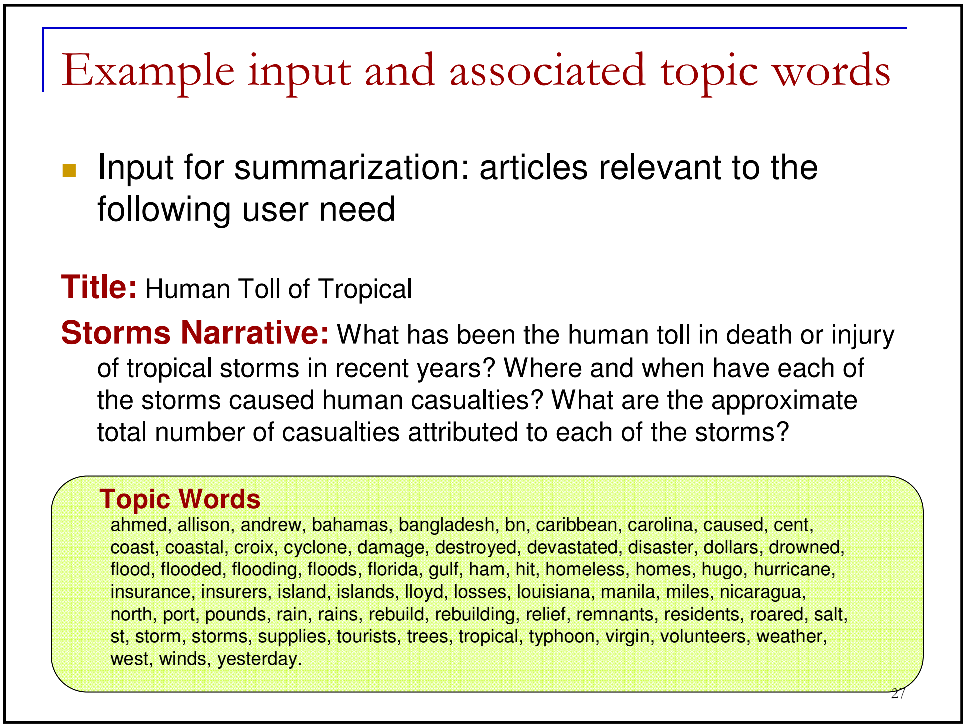
Global Optimization Methods — Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection

Speech Summarization — Segmentation, ASR
Acoustic Information, Disfluency

Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic

Topic words (topic signatures)

- Which words in the input are most descriptive?
  - Instead of assigning probabilities or weights to all words, divide words into two classes: descriptive or not
  - For iterative sentence selection approach, the binary distinction is key to the advantage over frequency and TF*IDF
  - Systems based on topic words have proven to be the most successful in official summarization evaluations



Example input and associated topic words

- Input for summarization: articles relevant to the following user need

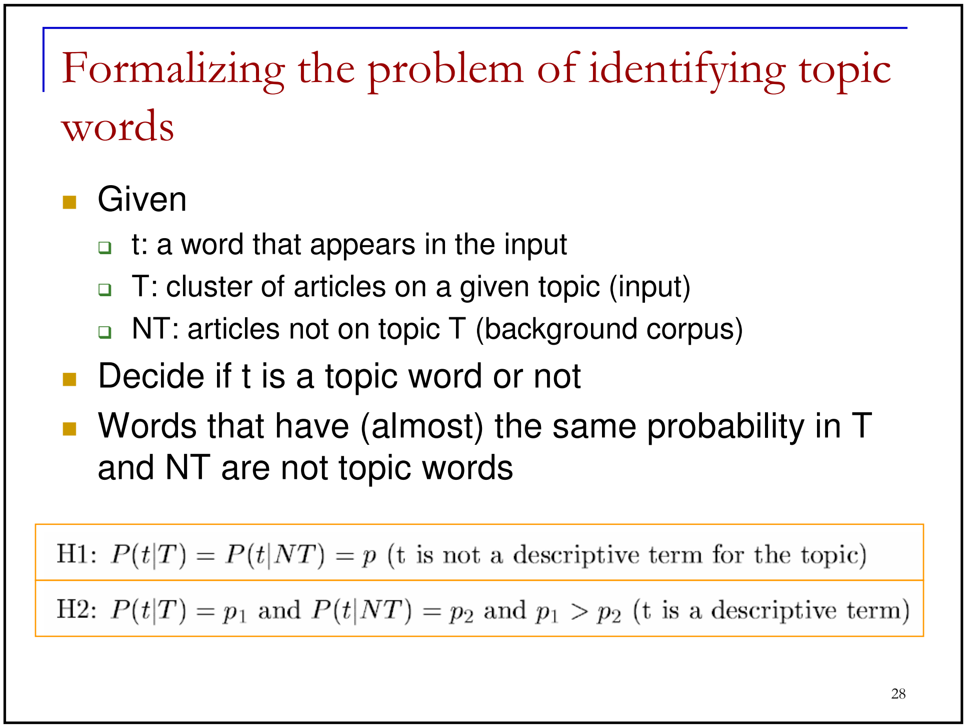
Title: Human Toll of Tropical

Storms Narrative: What has been the human toll in death or injury of tropical storms in recent years? Where and when have each of the storms caused human casualties? What are the approximate total number of casualties attributed to each of the storms?

Topic Words

ahmed, allison, andrew, bahamas, bangladesh, bn, caribbean, carolina, caused, cent, coast, coastal, croix, cyclone, damage, destroyed, devastated, disaster, dollars, drowned, flood, flooded, flooding, floods, florida, gulf, ham, hit, homeless, homes, hugo, hurricane, insurance, insurers, island, islands, lloyd, losses, louisiana, manila, miles, nicaragua, north, port, pounds, rain, rains, rebuild, rebuilding, relief, remnants, residents, roared, salt, st, storm, storms, supplies, tourists, trees, tropical, typhoon, virgin, volunteers, weather, west, winds, yesterday.

27



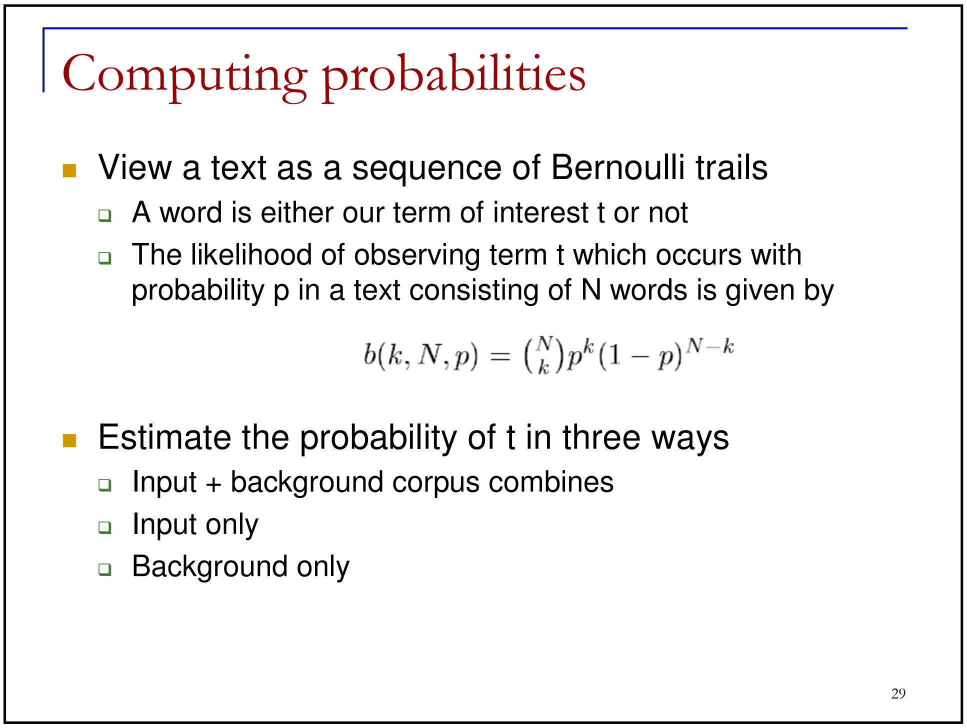
Formalizing the problem of identifying topic words

- Given
  - t: a word that appears in the input
  - T: cluster of articles on a given topic (input)
  - NT: articles not on topic T (background corpus)
- Decide if t is a topic word or not
- Words that have (almost) the same probability in T and NT are not topic words

H1: $P(t|T) = P(t|NT) = p$ (t is not a descriptive term for the topic)

H2: $P(t|T) = p_1$ and $P(t|NT) = p_2$ and $p_1 > p_2$ (t is a descriptive term)

28

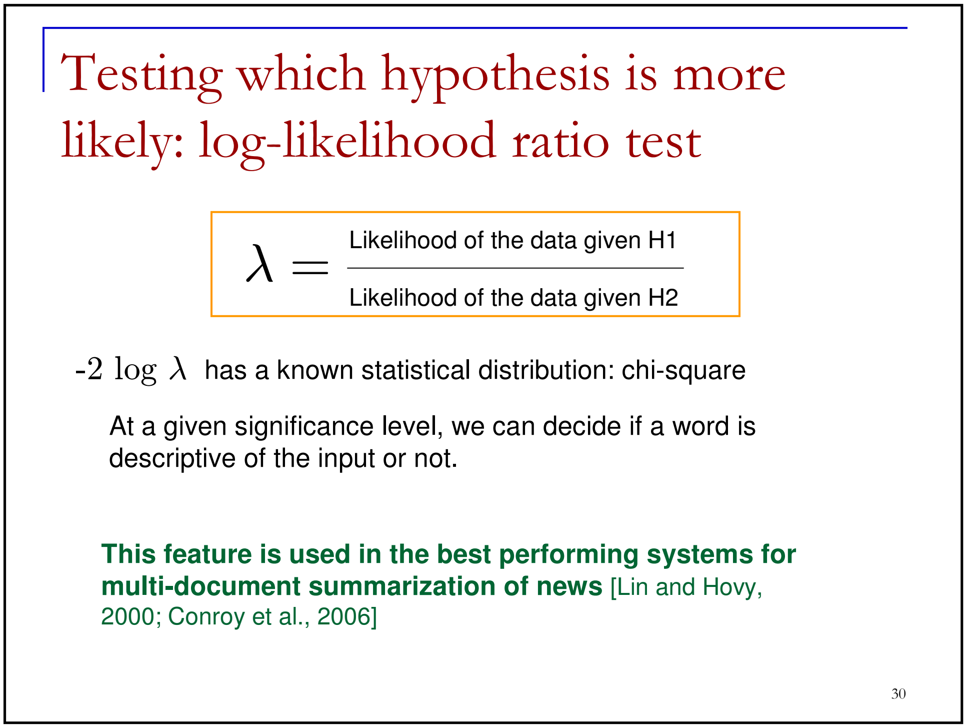


Computing probabilities

- View a text as a sequence of Bernoulli trails
  - A word is either our term of interest t or not
  - The likelihood of observing term t which occurs with probability p in a text consisting of N words is given by

$$b(k,N,p) = \binom{N}{k} p^k (1-p)^{N-k}$$

- Estimate the probability of t in three ways
  - Input + background corpus combines
  - Input only
  - Background only



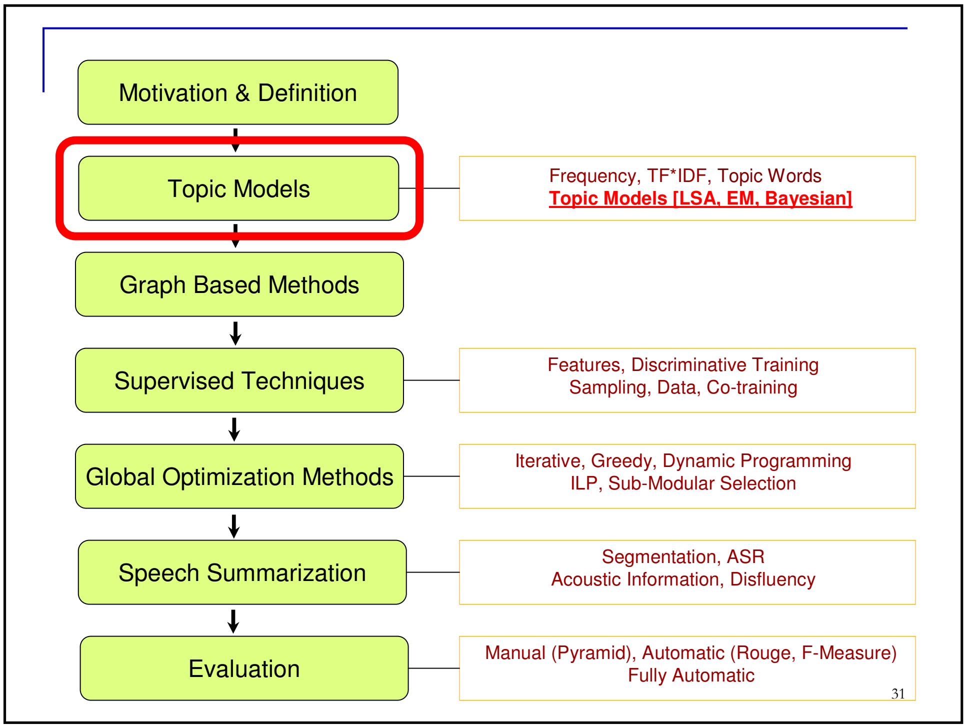
Testing which hypothesis is more likely: log-likelihood ratio test

$$\boxed{\lambda = \frac{\text{Likelihood of the data given H1}}{\text{Likelihood of the data given H2}}}$$

-2 $\log \lambda$ has a known statistical distribution: chi-square

At a given significance level, we can decide if a word is descriptive of the input or not.

This feature is used in the best performing systems for multi-document summarization of news [Lin and Hovy, 2000; Conroy et al., 2006]



Motivation & Definition

Topic Models — Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]

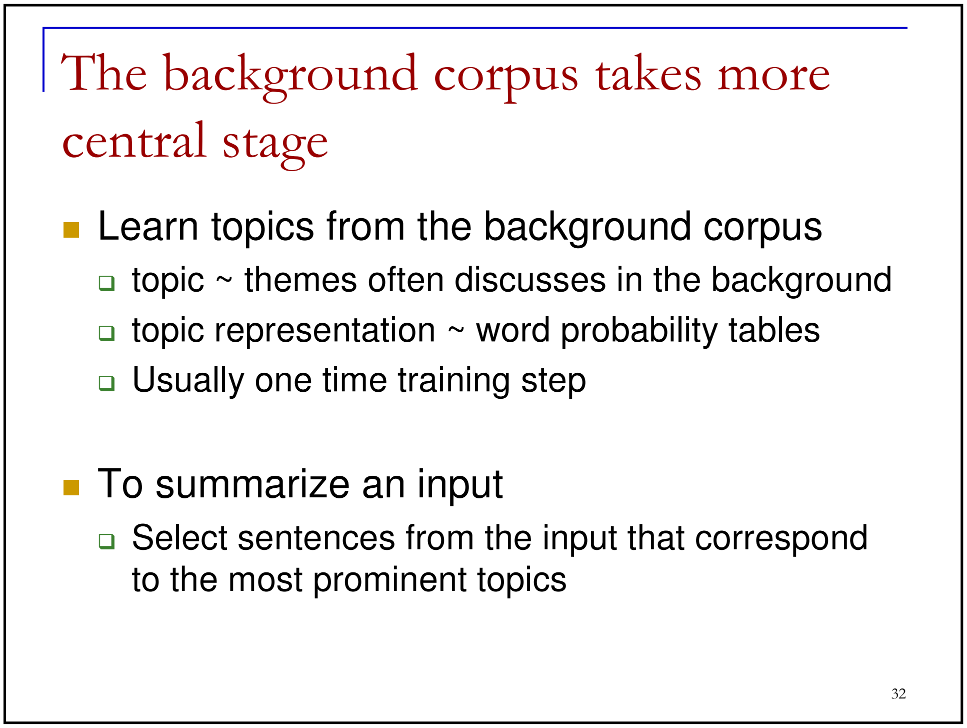
Graph Based Methods

Supervised Techniques — Features, Discriminative Training
Sampling, Data, Co-training

Global Optimization Methods — Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection

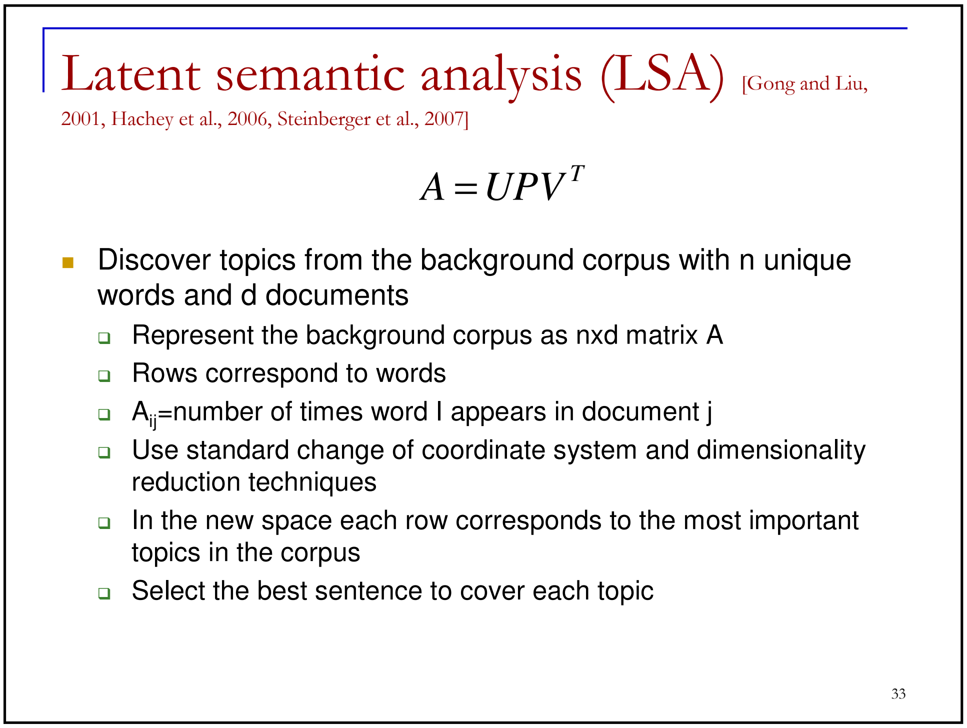
Speech Summarization — Segmentation, ASR
Acoustic Information, Disfluency

Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic



The background corpus takes more central stage

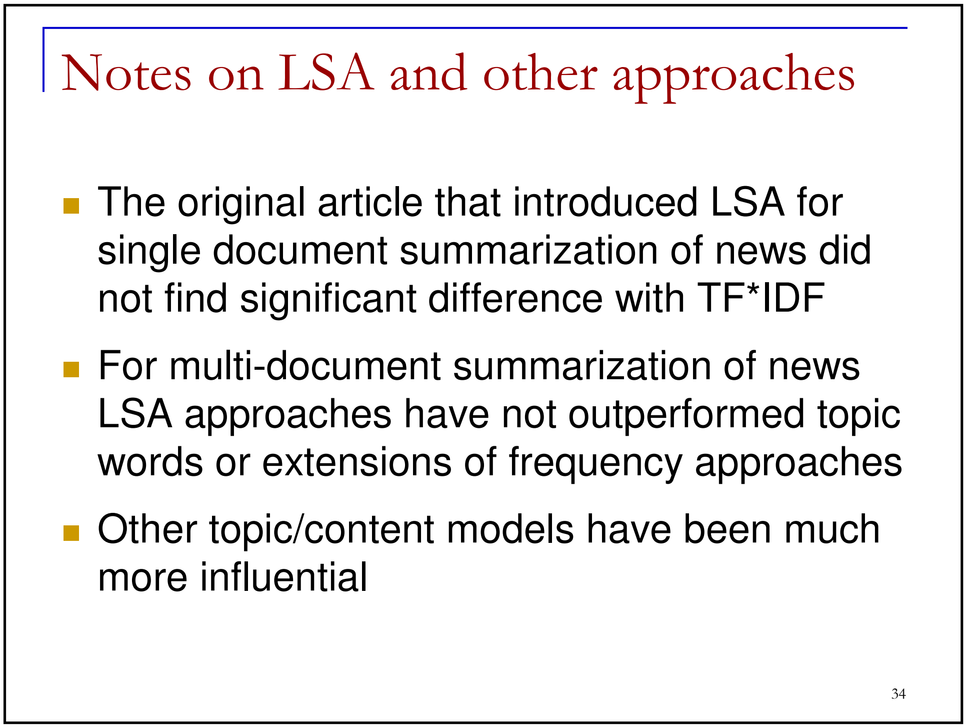
- Learn topics from the background corpus
  - topic ~ themes often discusses in the background
  - topic representation ~ word probability tables
  - Usually one time training step
- To summarize an input
  - Select sentences from the input that correspond to the most prominent topics



Latent semantic analysis (LSA) [Gong and Liu, 2001, Hachey et al., 2006, Steinberger et al., 2007]

$$A = UPV^T$$

- Discover topics from the background corpus with n unique words and d documents
  - Represent the background corpus as nxd matrix A
  - Rows correspond to words
  - A_{ij}=number of times word I appears in document j
  - Use standard change of coordinate system and dimensionality reduction techniques
  - In the new space each row corresponds to the most important topics in the corpus
  - Select the best sentence to cover each topic



Notes on LSA and other approaches

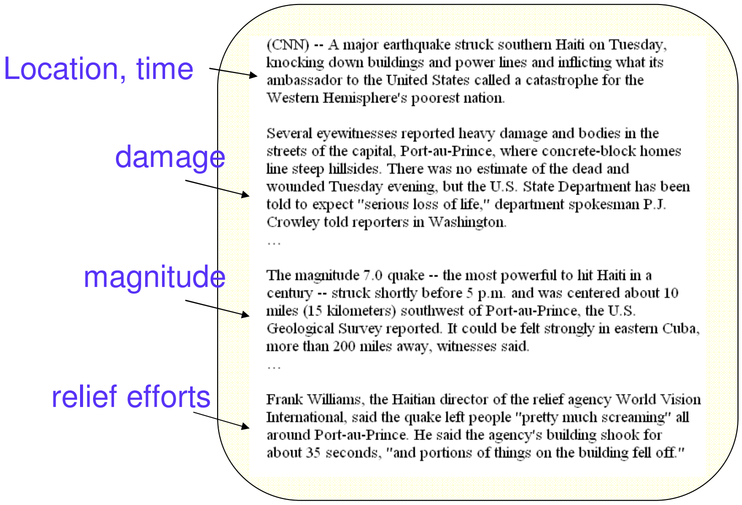
- The original article that introduced LSA for single document summarization of news did not find significant difference with TF*IDF
- For multi-document summarization of news LSA approaches have not outperformed topic words or extensions of frequency approaches
- Other topic/content models have been much more influential

Domain dependent content models

- Get sample documents from the domain
  - background corpus
- Cluster sentences from these documents
  - Implicit topics
- Obtain a word probability table for each topic
  - Counts only from the cluster representing the topic
- Select sentences from the input with highest probability for main topics

Text structure can be learnt

- Human-written examples from a domain



(CNN) -- A major earthquake struck southern Haiti on Tuesday, knocking down buildings and power lines and inflicting what its ambassador to the United States called a catastrophe for the Western Hemisphere's poorest nation.

Several eyewitnesses reported heavy damage and bodies in the streets of the capital, Port-au-Prince, where concrete-block homes line steep hillsides. There was no estimate of the dead and wounded Tuesday evening, but the U.S. State Department has been told to expect "serious loss of life," department spokesman P.J. Crowley told reporters in Washington.

…

The magnitude 7.0 quake -- the most powerful to hit Haiti in a century -- struck shortly before 5 p.m. and was centered about 10 miles (15 kilometers) southwest of Port-au-Prince, the U.S. Geological Survey reported. It could be felt strongly in eastern Cuba, more than 200 miles away, witnesses said.

…

Frank Williams, the Haitian director of the relief agency World Vision International, said the quake left people "pretty much screaming" all around Port-au-Prince. He said the agency's building shook for about 35 seconds, "and portions of things on the building fell off."

Topic = cluster of similar sentences from the background corpus

- Sentences cluster from earthquake articles
- Topic "earthquake location"
  - The Athens seismological institute said the temblor's epicenter was located 380 kilometers (238 miles) south of the capital.
  - Seismologists in Pakistan's Northwest Frontier Province said the temblor's epicenter was about 250 kilometers (155 miles) north of the provincial capital Peshawar.
  - The temblor was centered 60 kilometers (35 miles) north- west of the provincial capital of Kunming, about 2,200 kilometers (1,300 miles) southwest of Beijing, a bureau seismologist said.

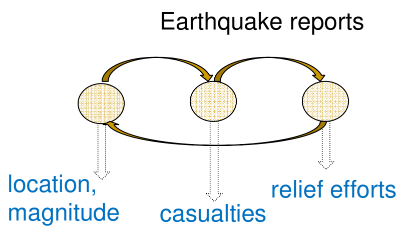
Content model [Barzilay and Lee, 2004, Pascale et al., 2003]

- Hidden Markov Model (HMM)-based
  - States - clusters of related sentences "topics"
  - Transition prob. - sentence precedence in corpus
  - Emission prob. - bigram language model

Generating sentence in current *topic*

$$p(<s_{i+1},h_{i+1}>|<s_i,h_i>) = p_t(h_{i+1}\,|\,h_i)\cdot p_e(s_{i+1}\,|\,h_{i+1})$$

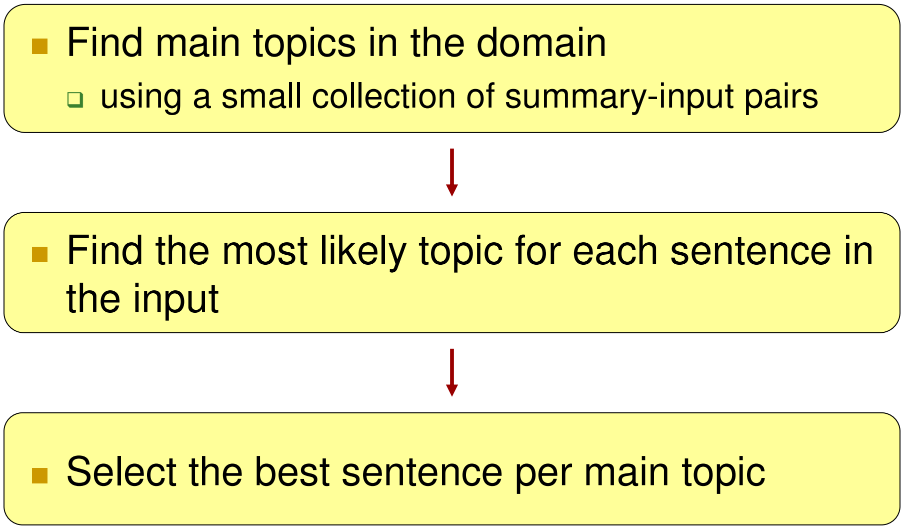
Transition from previous *topic*



Learning the content model

- Many articles from the same domain
- Cluster sentences: each cluster represents a topic from the domain
  - Word probability tables for each topic
- Transitions between clusters can be computed from sentence adjacencies in the original articles
  - Probabilities of going from one topic to another
- Iterate between clustering and transition probability estimation to obtain domain model

To select a summary



- Find main topics in the domain
  - using a small collection of summary-input pairs
- Find the most likely topic for each sentence in the input
- Select the best sentence per main topic

Historical note

- Some early approaches to multi-document summarization relied on clustering the sentences in the input alone [McKeown et al., 1999, Siddharthan et al., 2004]
  - Clusters of similar sentences represent a theme in the input
  - Clusters with more sentences are more important
  - Select one sentence per important cluster

Example cluster

Choose one sentence to represent the cluster

1. PAL was devastated by a pilots' strike in June and by the region's currency crisis.

2. In June, PAL was embroiled in a crippling three-week pilots' strike.

3. Tan wants to retain the 200 pilots because they stood by him when the majority of PAL's pilots staged a devastating strike in June.

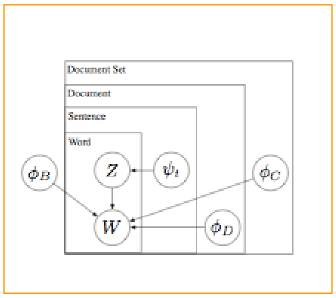
Bayesian content models

- Takes a batch of inputs for summarization
- Many word probability tables
  - One for general English
  - One for each of the inputs to be summarized
  - One for each document in any input

To select a summary S with L words from document collection D given as input

$$S^* = \min_{S:words(S)\leq L} KL(P_D||P_S)$$

The goal is to select the summary, not a sentence. Greedy selection vs. global will be discussed in detail later



KL divergence

- Distance between two probability distributions: P, Q

$$KL(P \| Q) = \sum_w p_P(w) \log_2 \frac{p_P(w)}{p_Q(w)}$$

- P, Q: Input and summary word distributions

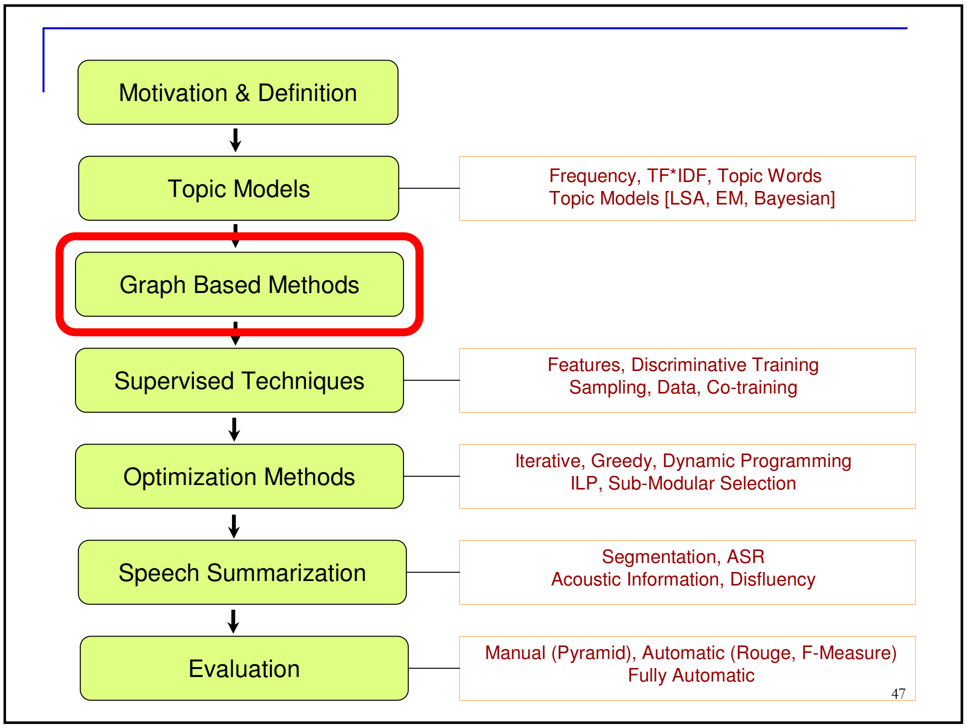
Intriguing side note

- In the full Bayesian topic models, word probabilities for all words is more important than binary distinctions of topic and non-topic word

- Haghighi and Vanderwende report that a system that chooses the summary with highest expected number of topic words performs as SumBasic

Review

- Frequency based informativeness has been used in building summarizers
- Topic words probably more useful
- Topic models
  - Latent Semantic Analysis
  - Domain dependent content model
  - Bayesian content model



Motivation & Definition

Topic Models — Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]

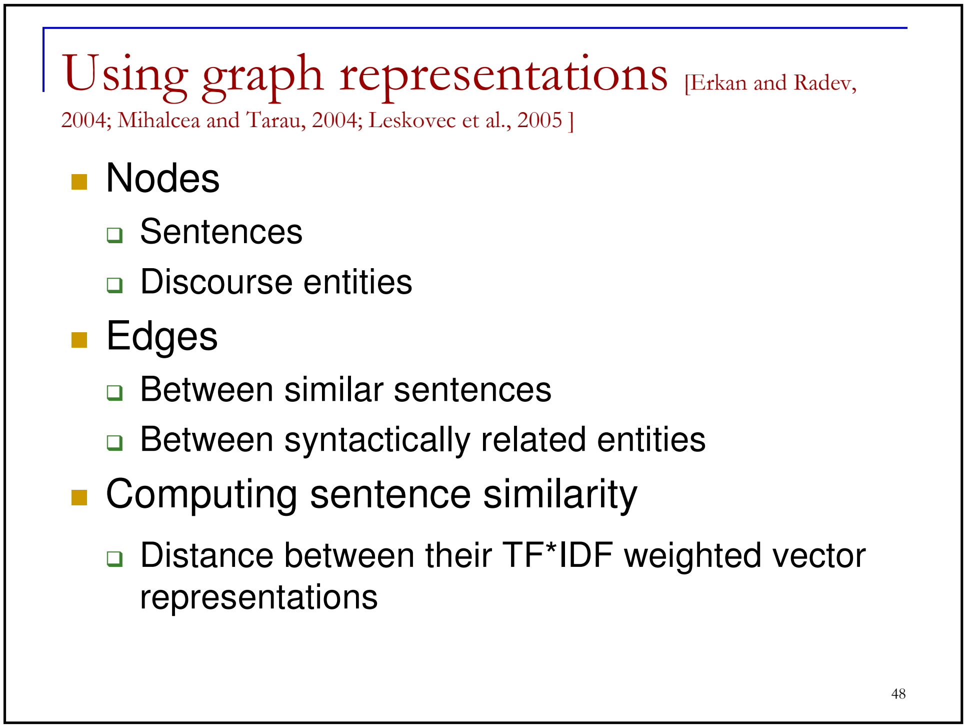
Graph Based Methods

Supervised Techniques — Features, Discriminative Training
Sampling, Data, Co-training

Optimization Methods — Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection

Speech Summarization — Segmentation, ASR
Acoustic Information, Disfluency

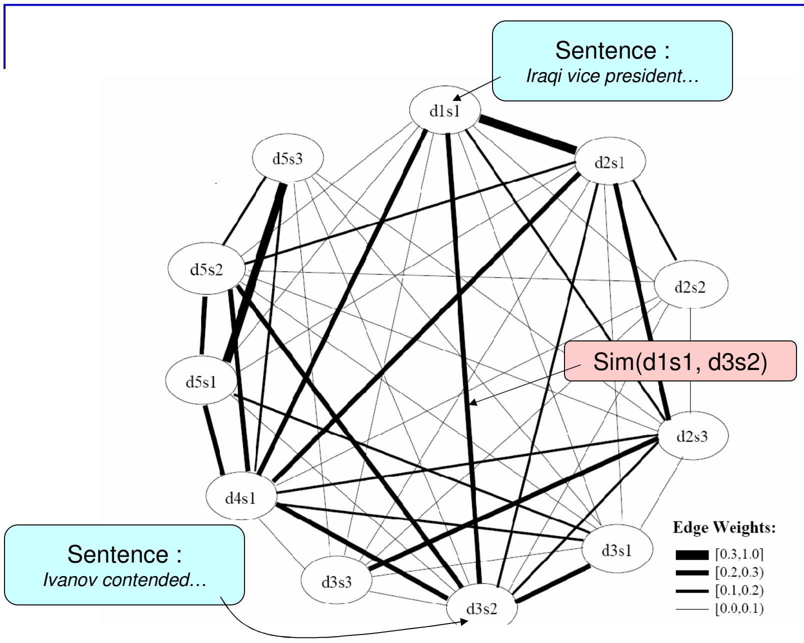
Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic



Using graph representations [Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Leskovec et al., 2005]

- Nodes
  - Sentences
  - Discourse entities
- Edges
  - Between similar sentences
  - Between syntactically related entities
- Computing sentence similarity
  - Distance between their TF*IDF weighted vector representations

SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi



Sentence :
Iraqi vice president…

Sim(d1s1, d3s2)

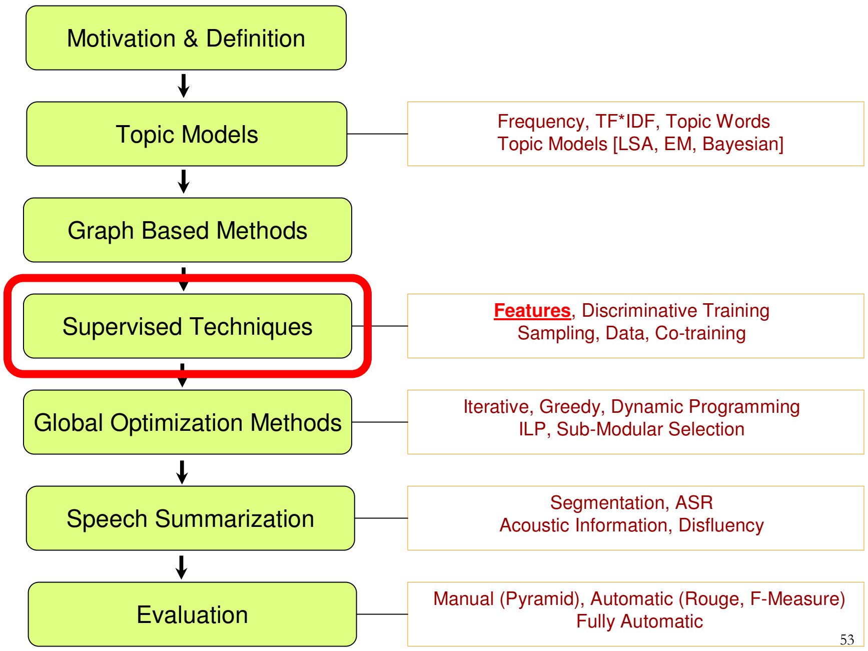
Sentence :
Ivanov contended…

Advantages of the graph model

- Combines word frequency and sentence clustering
- Gives a formal model for computing importance: random walks
  - Normalize weights of edges to sum to 1
  - They now represent probabilities of transitioning from one node to another

Random walks for summarization

- Represent the input text as graph
- Start traversing from node to node
  - following the transition probabilities
  - occasionally hopping to a new node
- What is the probability that you are in any particular node after doing this process for a certain time?
  - Standard solution (stationary distribution)
  - This probability is the weight of the sentence



Motivation & Definition

Topic Models

Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]

Graph Based Methods

Supervised Techniques

Features, Discriminative Training
Sampling, Data, Co-training

Global Optimization Methods

Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection

Speech Summarization

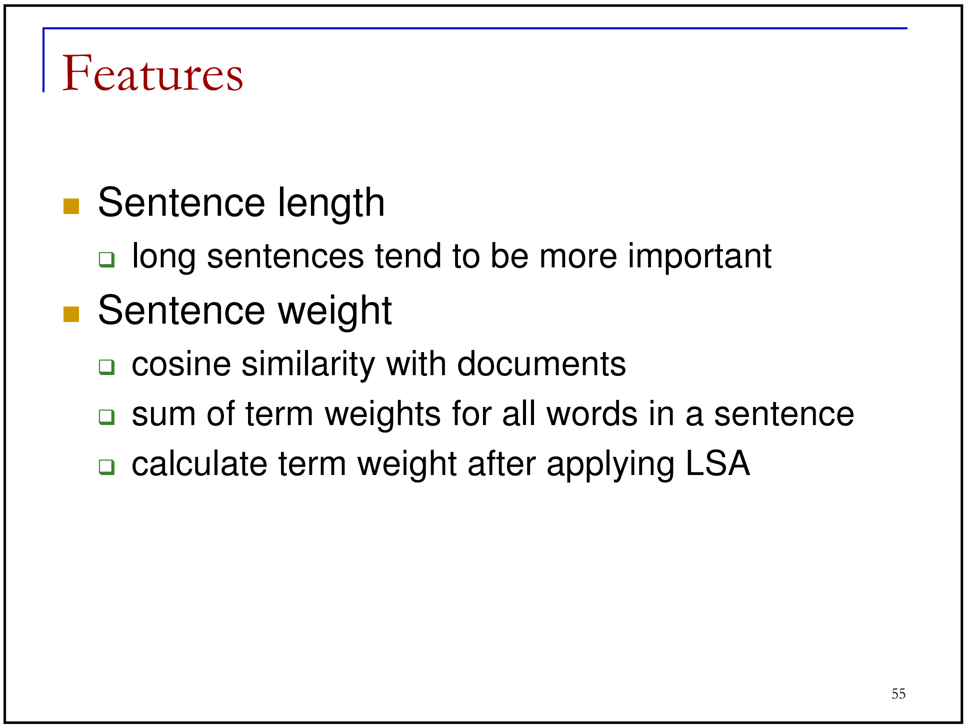
Segmentation, ASR
Acoustic Information, Disfluency

Evaluation

Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic

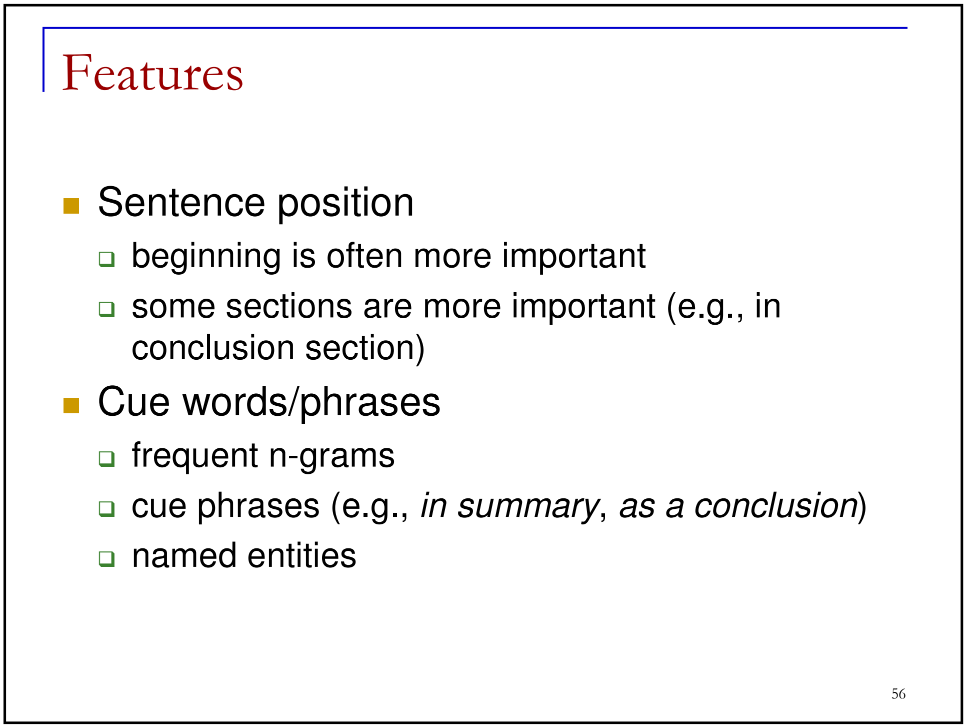
Supervised methods

- For extractive summarization, the task can be represented as binary classification
  - A sentence is in the summary or not
- Use statistical classifiers to determine the score of a sentence: how likely it’s included in the summary
  - Feature representation for each sentence
  - Classification models trained from annotated data
- Select the sentences with highest scores (greedy for now, see other selection methods later)



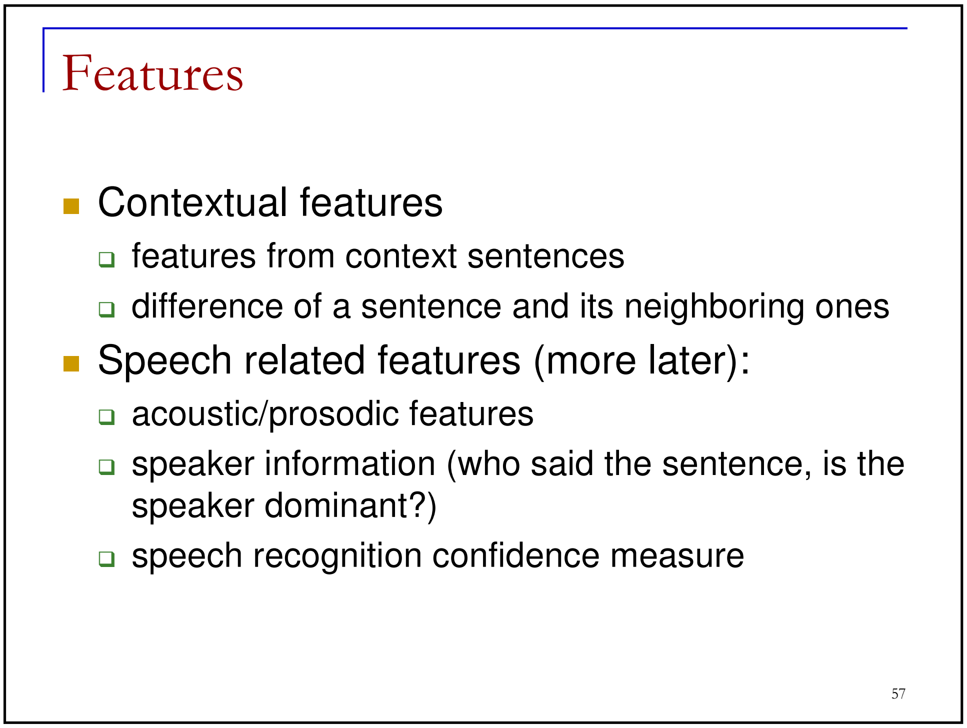
Features

- Sentence length
  - long sentences tend to be more important
- Sentence weight
  - cosine similarity with documents
  - sum of term weights for all words in a sentence
  - calculate term weight after applying LSA



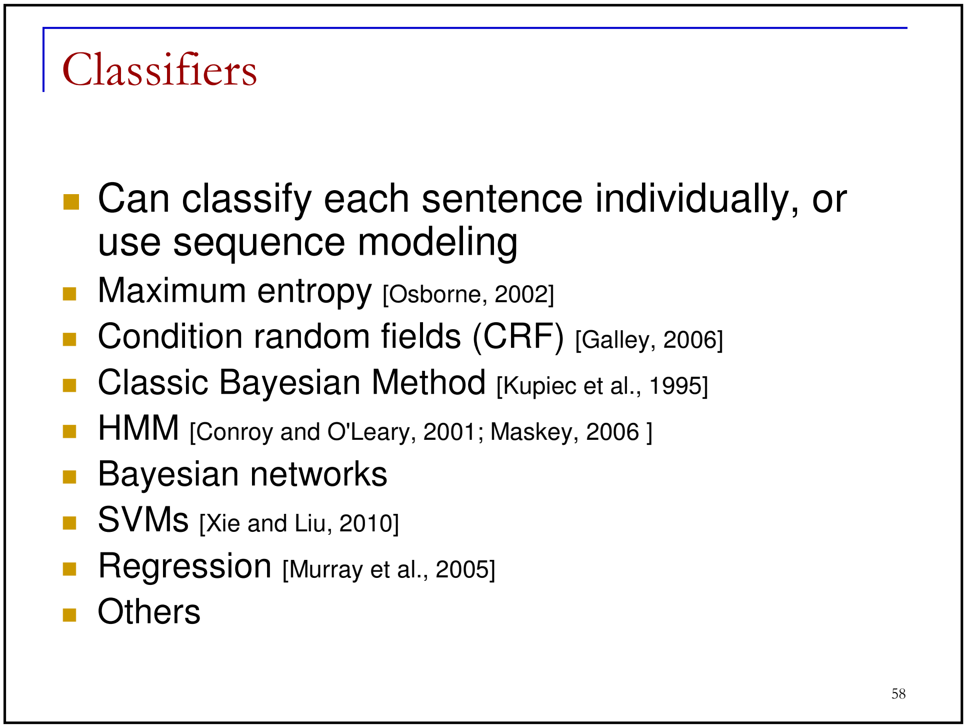
Features

- Sentence position
  - beginning is often more important
  - some sections are more important (e.g., in conclusion section)
- Cue words/phrases
  - frequent n-grams
  - cue phrases (e.g., *in summary*, *as a conclusion*)
  - named entities



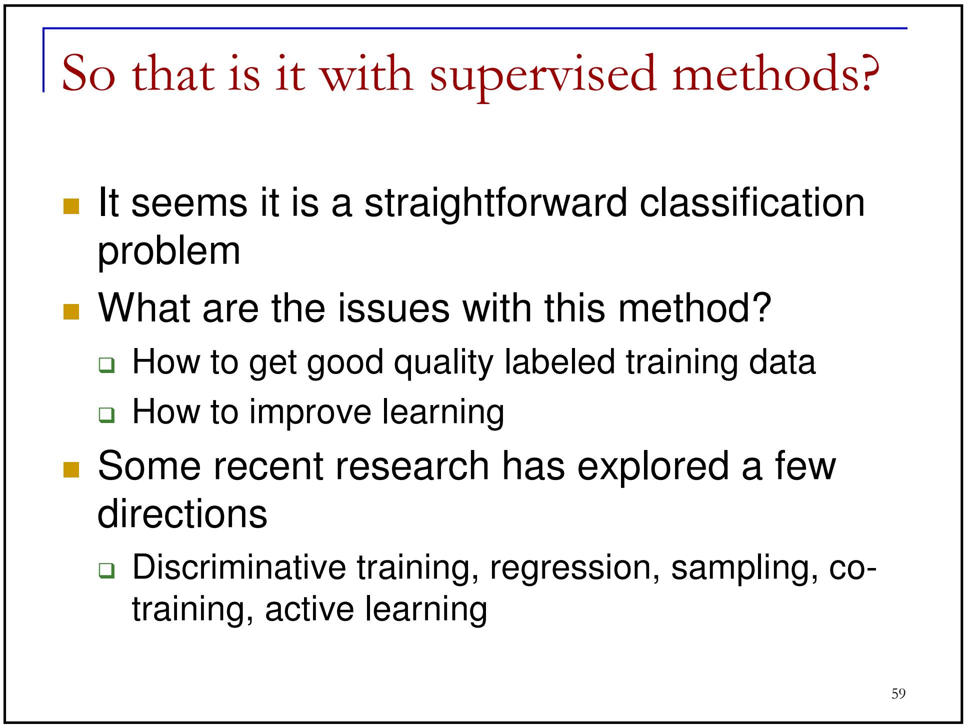
Features

- Contextual features
  - features from context sentences
  - difference of a sentence and its neighboring ones
- Speech related features (more later):
  - acoustic/prosodic features
  - speaker information (who said the sentence, is the speaker dominant?)
  - speech recognition confidence measure



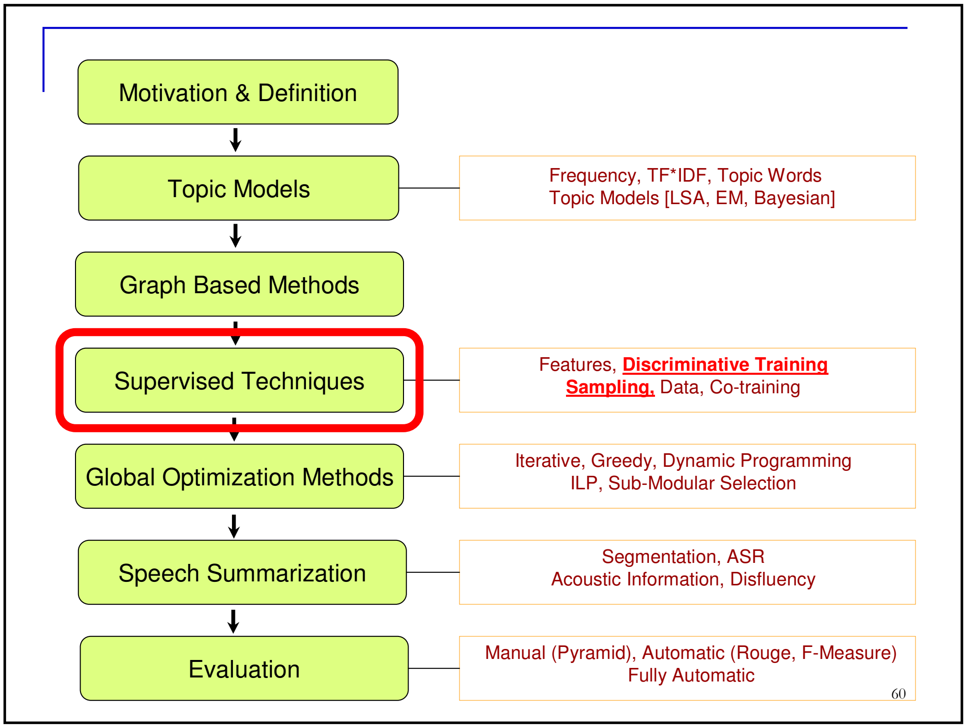
Classifiers

- Can classify each sentence individually, or use sequence modeling
- Maximum entropy [Osborne, 2002]
- Condition random fields (CRF) [Galley, 2006]
- Classic Bayesian Method [Kupiec et al., 1995]
- HMM [Conroy and O'Leary, 2001; Maskey, 2006]
- Bayesian networks
- SVMs [Xie and Liu, 2010]
- Regression [Murray et al., 2005]
- Others



So that is it with supervised methods?

- It seems it is a straightforward classification problem
- What are the issues with this method?
  - How to get good quality labeled training data
  - How to improve learning
- Some recent research has explored a few directions
  - Discriminative training, regression, sampling, co-training, active learning



Motivation & Definition

Topic Models — Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]

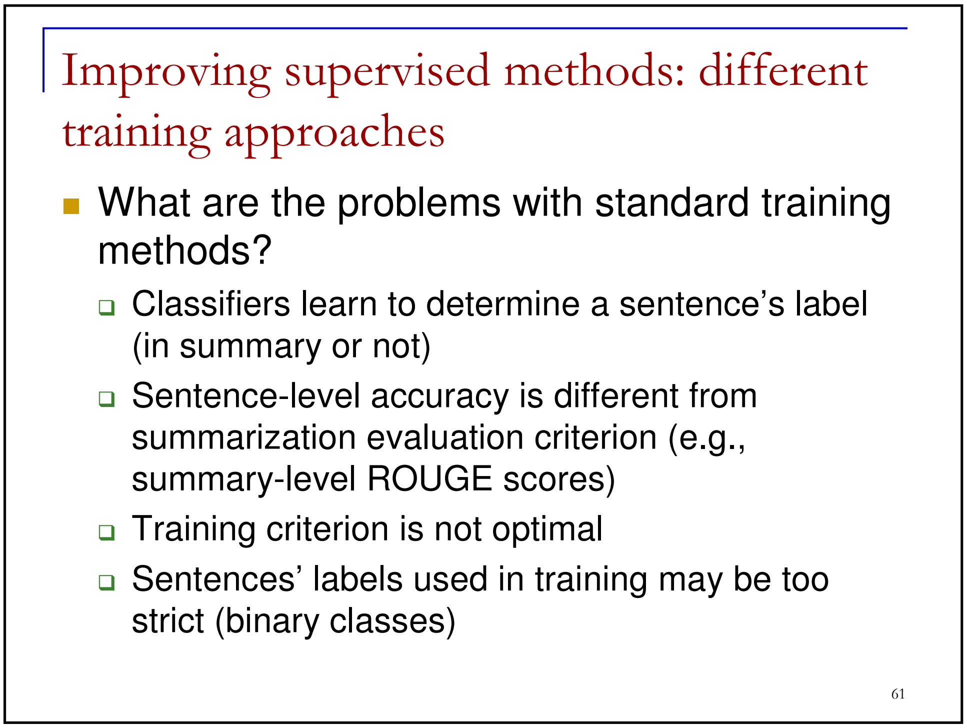
Graph Based Methods

Supervised Techniques — Features, **Discriminative Training** **Sampling,** Data, Co-training

Global Optimization Methods — Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection

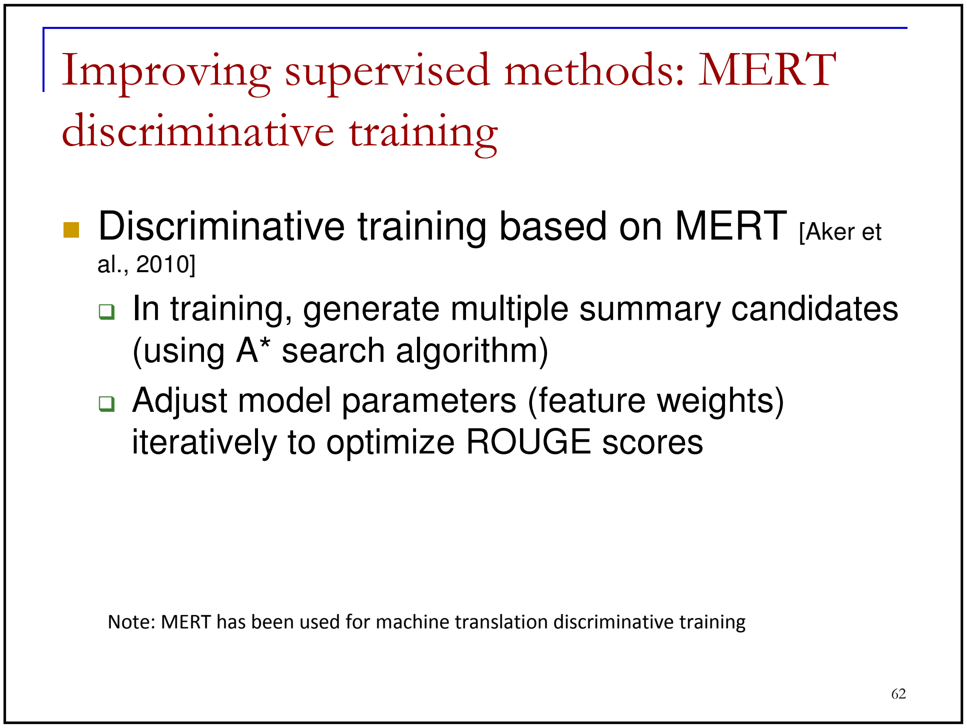
Speech Summarization — Segmentation, ASR
Acoustic Information, Disfluency

Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic



Improving supervised methods: different training approaches

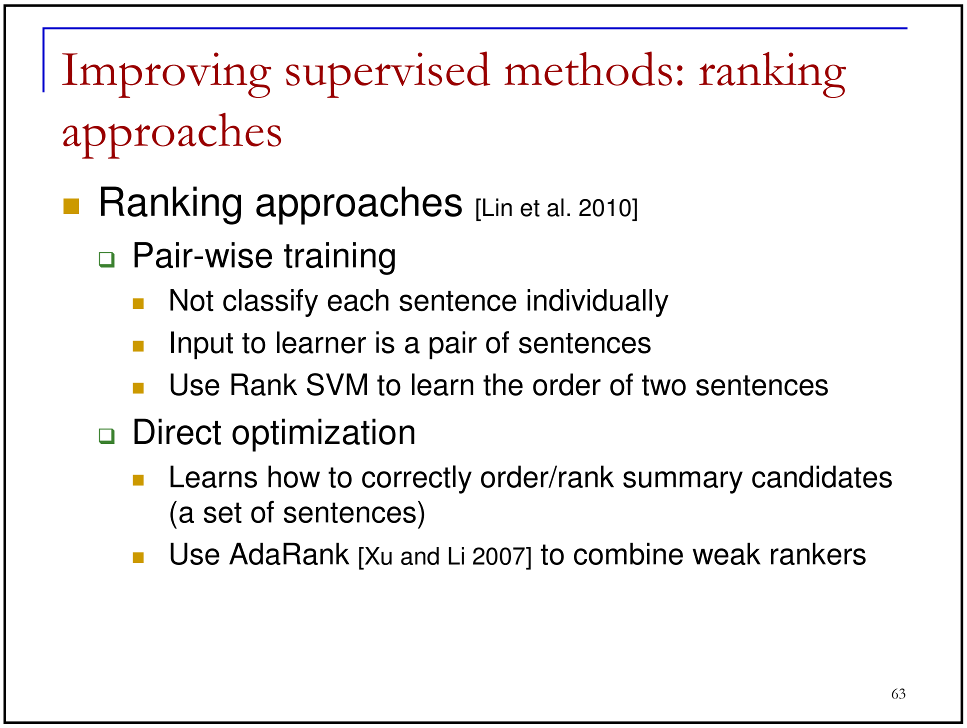
- What are the problems with standard training methods?
  - Classifiers learn to determine a sentence’s label (in summary or not)
  - Sentence-level accuracy is different from summarization evaluation criterion (e.g., summary-level ROUGE scores)
  - Training criterion is not optimal
  - Sentences’ labels used in training may be too strict (binary classes)



Improving supervised methods: MERT discriminative training

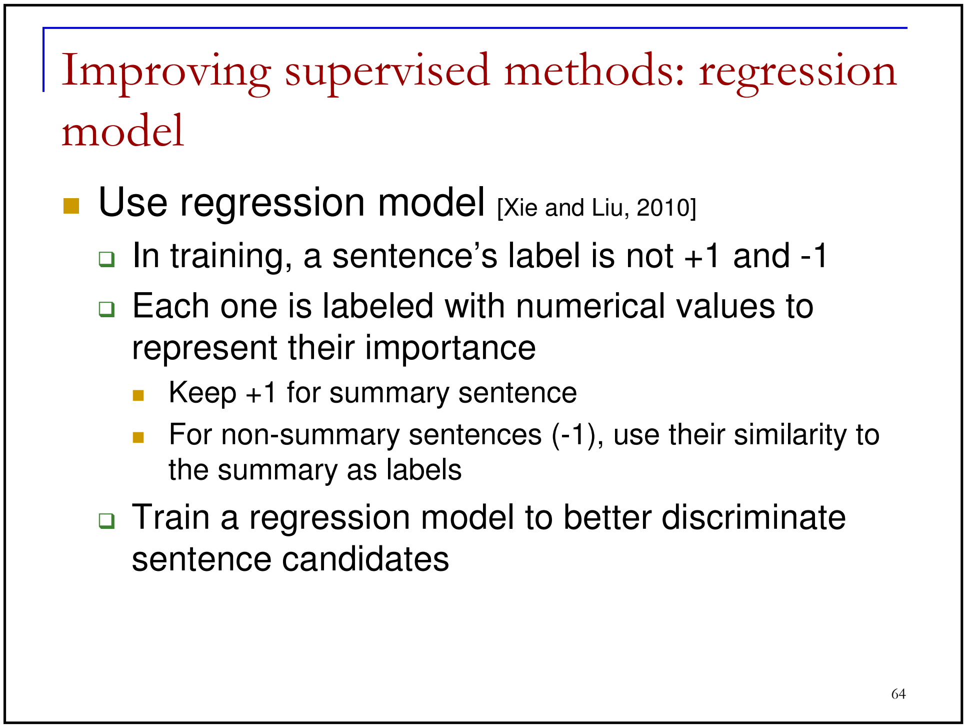
- Discriminative training based on MERT [Aker et al., 2010]
  - In training, generate multiple summary candidates (using A* search algorithm)
  - Adjust model parameters (feature weights) iteratively to optimize ROUGE scores

Note: MERT has been used for machine translation discriminative training



Improving supervised methods: ranking approaches

- Ranking approaches [Lin et al. 2010]
  - Pair-wise training
    - Not classify each sentence individually
    - Input to learner is a pair of sentences
    - Use Rank SVM to learn the order of two sentences
  - Direct optimization
    - Learns how to correctly order/rank summary candidates (a set of sentences)
    - Use AdaRank [Xu and Li 2007] to combine weak rankers



Improving supervised methods: regression model

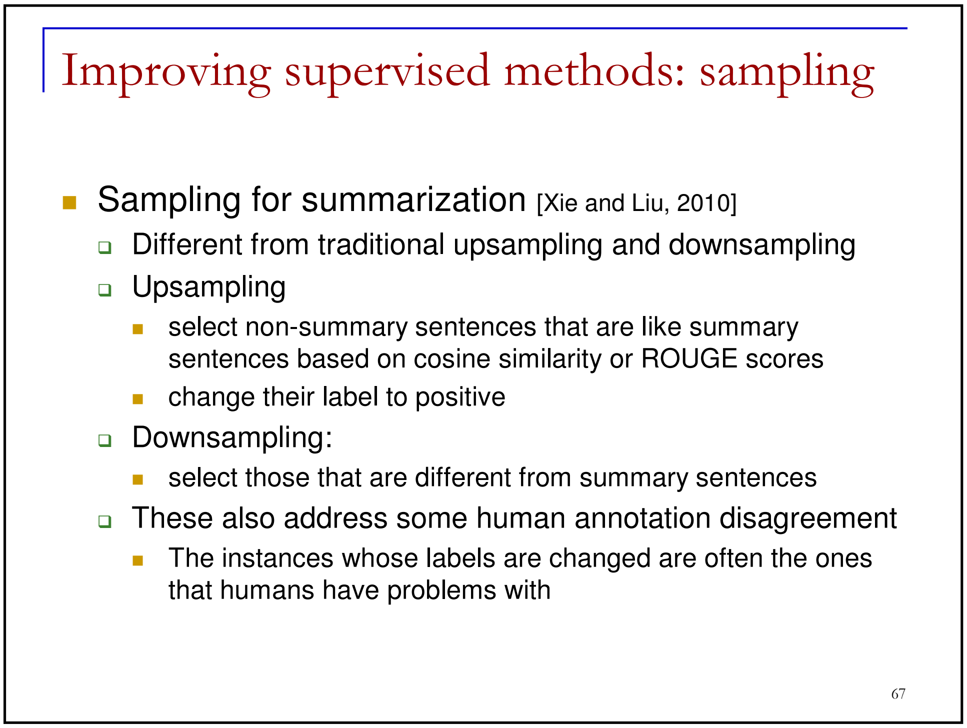
- Use regression model [Xie and Liu, 2010]
  - In training, a sentence’s label is not +1 and -1
  - Each one is labeled with numerical values to represent their importance
    - Keep +1 for summary sentence
    - For non-summary sentences (-1), use their similarity to the summary as labels
  - Train a regression model to better discriminate sentence candidates

Improving supervised methods: sampling

- **Problems** -- in binary classification setup for summarization, the two classes are imbalanced
  - Summary sentences are minority class.
  - Imbalanced data can hurt classifier training
- How can we address this?
  - Sampling to make distribution more balanced to train classifiers
  - Has been studied a lot in machine learning

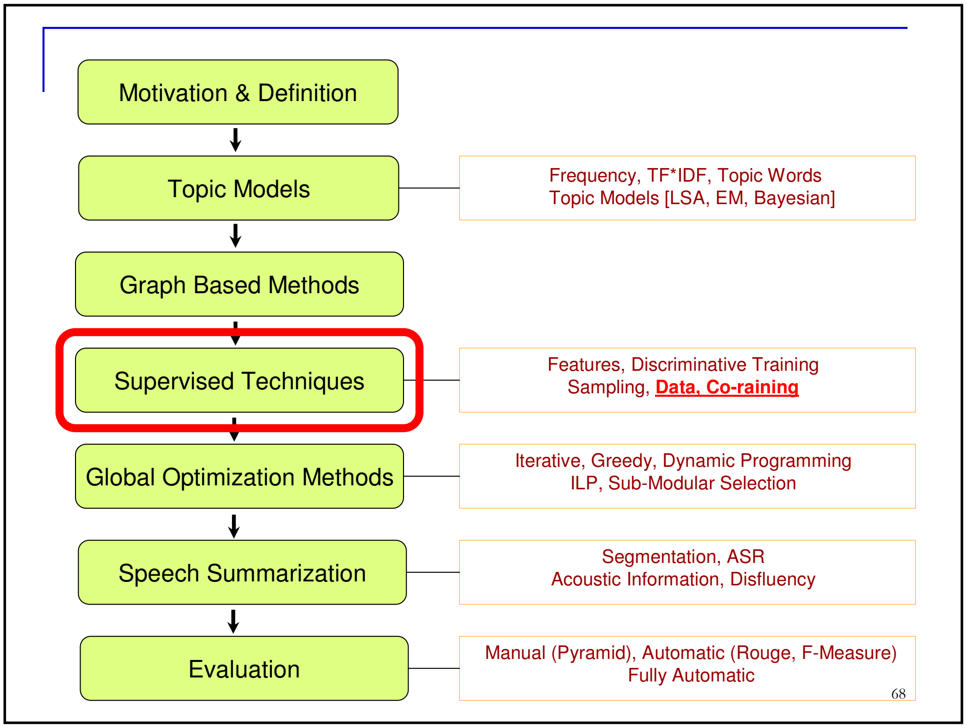
Improving supervised methods: sampling

- Upsampling: increase minority samples
  - Replicate existing minority samples
  - Generate synthetic examples (e.g., by some kind of interpolation)
- Downsampling: reduce majority samples
  - Often randomly select from existing majority samples



Improving supervised methods: sampling

- Sampling for summarization [Xie and Liu, 2010]
  - Different from traditional upsampling and downsampling
  - Upsampling
    - select non-summary sentences that are like summary sentences based on cosine similarity or ROUGE scores
    - change their label to positive
  - Downsampling:
    - select those that are different from summary sentences
  - These also address some human annotation disagreement
    - The instances whose labels are changed are often the ones that humans have problems with



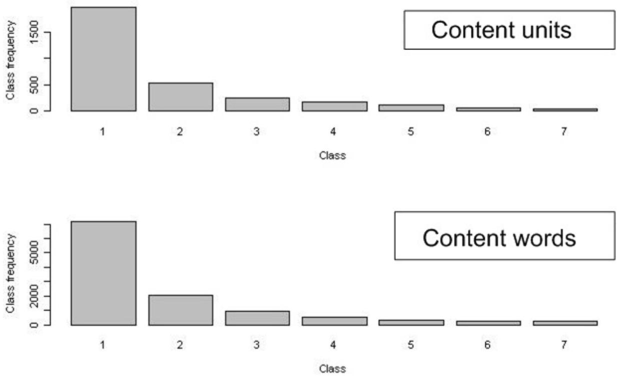
Motivation & Definition
Topic Models
Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]
Graph Based Methods
Supervised Techniques
Features, Discriminative Training
Sampling, Data, Co-raining
Global Optimization Methods
Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection
Speech Summarization
Segmentation, ASR
Acoustic Information, Disfluency
Evaluation
Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic

Supervised methods: data issues

- Need labeled data for model training
- How do we get good quality training data?
  - Can ask human annotators to select extractive summary sentences
  - However, human agreement is generally low
- What if data is not labeled at all? or it only has abstractive summary?

Do humans agree on summary sentence selection?

Human agreement on word/sentence/fact selection



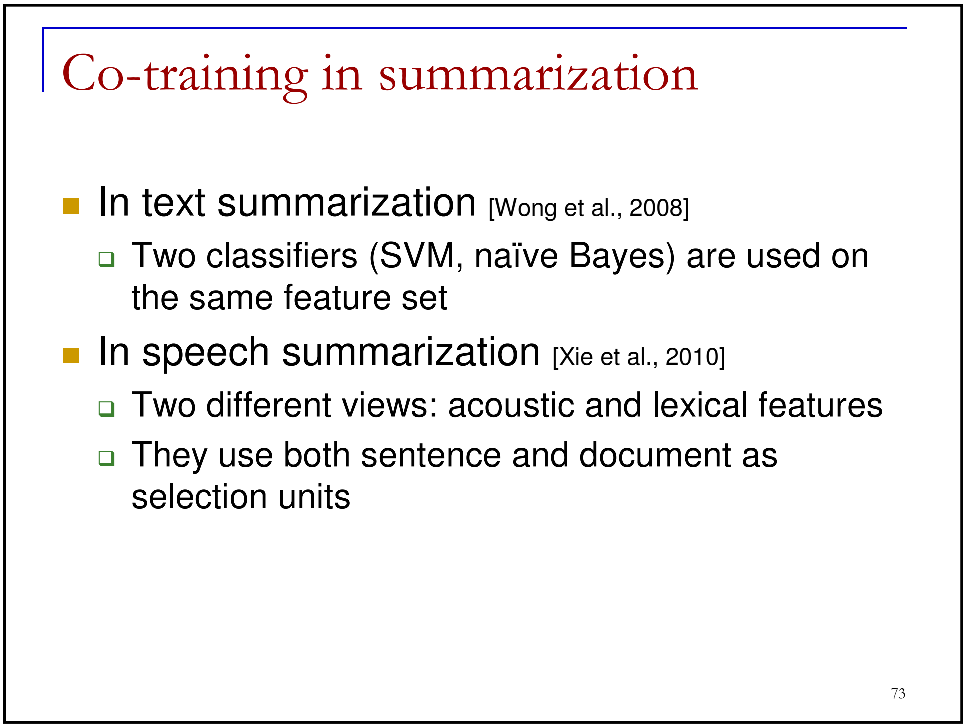
- Distributions of content units and words are similar
- Few units are expressed by everyone; many units are expressed by only one person

Supervised methods: semi-supervised learning

- **Question** – can we use unlabeled data to help supervised methods?
- A lot of research has been done on semi-supervised learning for various tasks
- Co-training and active learning have been used in summarization

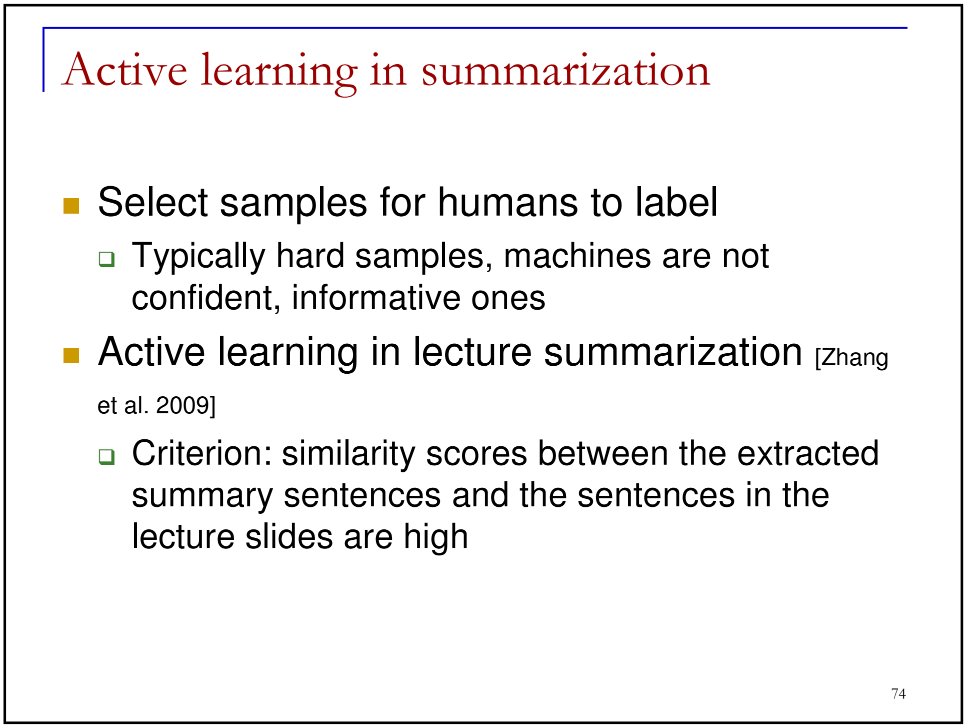
Co-training

- Use co-training to leverage unlabeled data
  - Feature sets represent different views
  - They are conditionally independent given the class label
  - Each is sufficient for learning
  - Select instances based on one view, to help the other classifier



Co-training in summarization

- In text summarization [Wong et al., 2008]
  - Two classifiers (SVM, naïve Bayes) are used on the same feature set
- In speech summarization [Xie et al., 2010]
  - Two different views: acoustic and lexical features
  - They use both sentence and document as selection units



Active learning in summarization

- Select samples for humans to label
  - Typically hard samples, machines are not confident, informative ones
- Active learning in lecture summarization [Zhang et al. 2009]
  - Criterion: similarity scores between the extracted summary sentences and the sentences in the lecture slides are high

Supervised methods: using labeled abstractive summaries

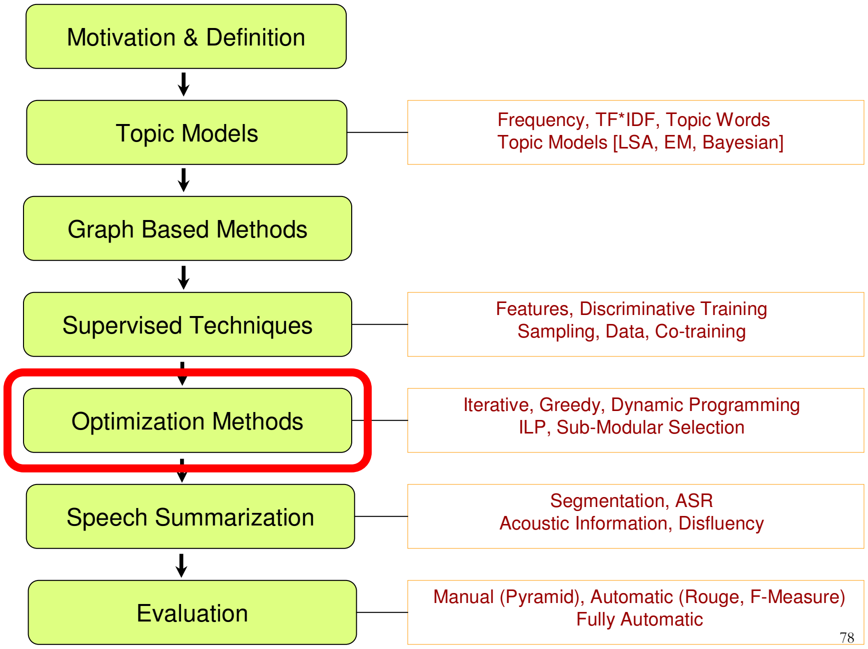
- **Question** -- what if I only have abstractive summaries, but not extractive summaries?
- No labeled sentences to use for classifier training in extractive summarization
- Can use reference abstract summary to automatically create labels for sentences
  - Use similarity of a sentence to the human written abstract (or ROUGE scores, other metrics)

Comment on supervised performance

- Easier to incorporate more information
- At the cost of requiring a large set of human annotated training data
- Human agreement is low, therefore labeled training data is noisy
- Need matched training/test conditions
  - may not easily generalize to different domains
- Effective features vary for different domains
  - e.g., position is important for news articles

Comments on supervised performance

- Seems supervised methods are more successful in speech summarization than in text
  - Speech summarization is almost never multi-document
  - There are fewer indications about the topic of the input in speech domains
  - Text analysis techniques used in speech summarization are relatively simpler



Motivation & Definition

Topic Models — Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]

Graph Based Methods

Supervised Techniques — Features, Discriminative Training
Sampling, Data, Co-training

Optimization Methods — Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection

Speech Summarization — Segmentation, ASR
Acoustic Information, Disfluency

Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic

Parameters to optimize

- In summarization methods we try to find

1. Most significant sentences
2. Remove redundant ones
3. Keep the summary under given length

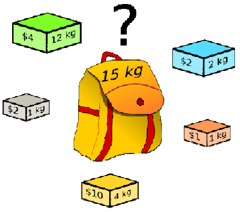
- ❑ Can we combine all 3 steps in one?
  - ❑ Optimize all 3 parameters at once

Summarization as an optimization problem

- Summarization Problem

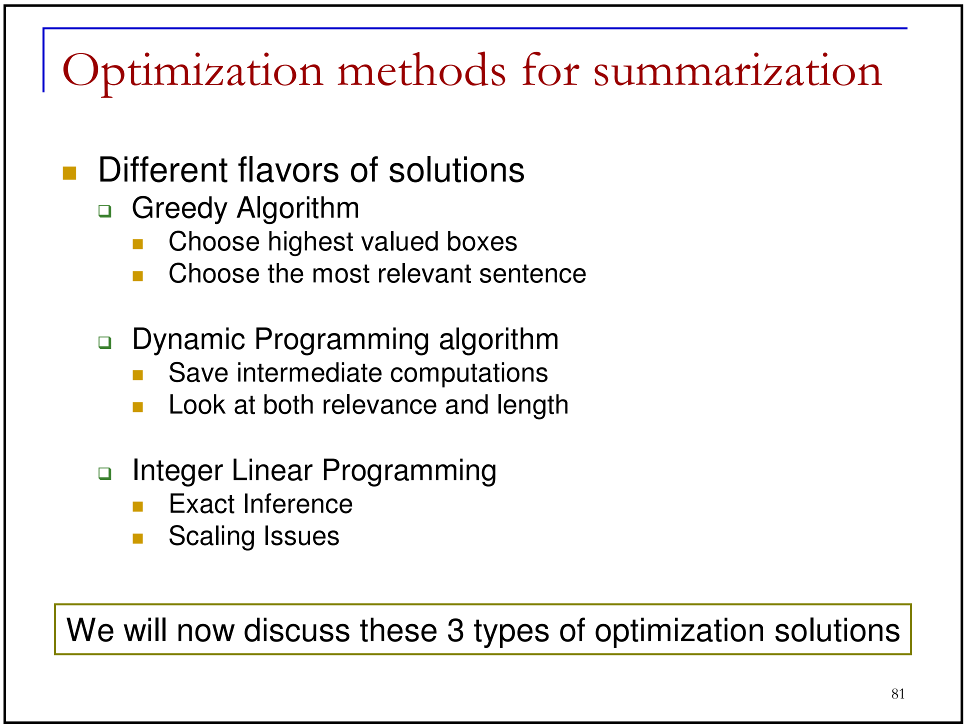
Select sentences such that summary relevance is maximized while keeping total length under X words

- Knapsack Optimization Problem



Select boxes such that amount of money is maximized while keeping total weight under X Kg

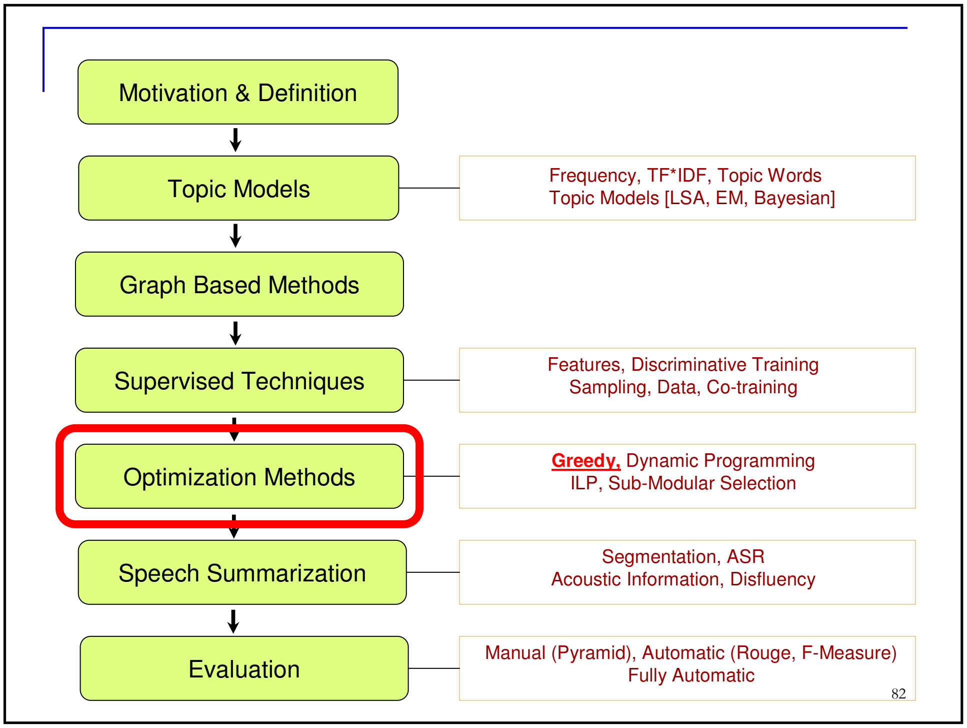
- Many other similar optimization problems
- General Idea: Maximize a function given a set of constraints



Optimization methods for summarization

- Different flavors of solutions
  - Greedy Algorithm
    - Choose highest valued boxes
    - Choose the most relevant sentence
  - Dynamic Programming algorithm
    - Save intermediate computations
    - Look at both relevance and length
  - Integer Linear Programming
    - Exact Inference
    - Scaling Issues

We will now discuss these 3 types of optimization solutions



Motivation & Definition

Topic Models — Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]

Graph Based Methods

Supervised Techniques — Features, Discriminative Training
Sampling, Data, Co-training

Optimization Methods — **Greedy,** Dynamic Programming
ILP, Sub-Modular Selection

Speech Summarization — Segmentation, ASR
Acoustic Information, Disfluency

Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic

Greedy optimization algorithms

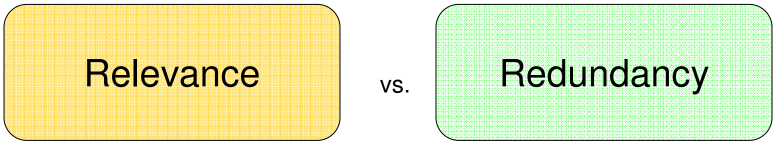
- Greedy solution is an approximate algorithm which may not be optimal

- Choose the most relevant + least redundant sentence if the total length does not exceed the summary length
  - Maximal Marginal Relevance is one such greedy algorithm proposed by [Carbonell et al., 1998]

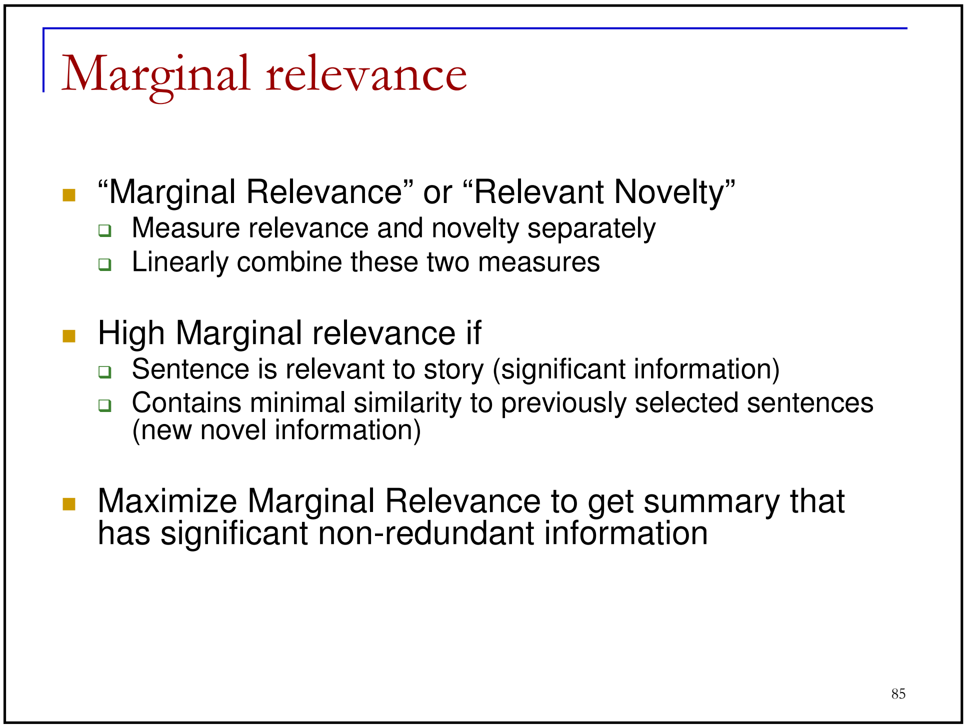
Maximal Marginal Relevance (MMR)

[Carbonell et al., 1998]

- Summary: relevant and non-redundant information
  - Many summaries are built based on sentences ranked by relevance
  - E.g. Extract most relevant 30% of sentences

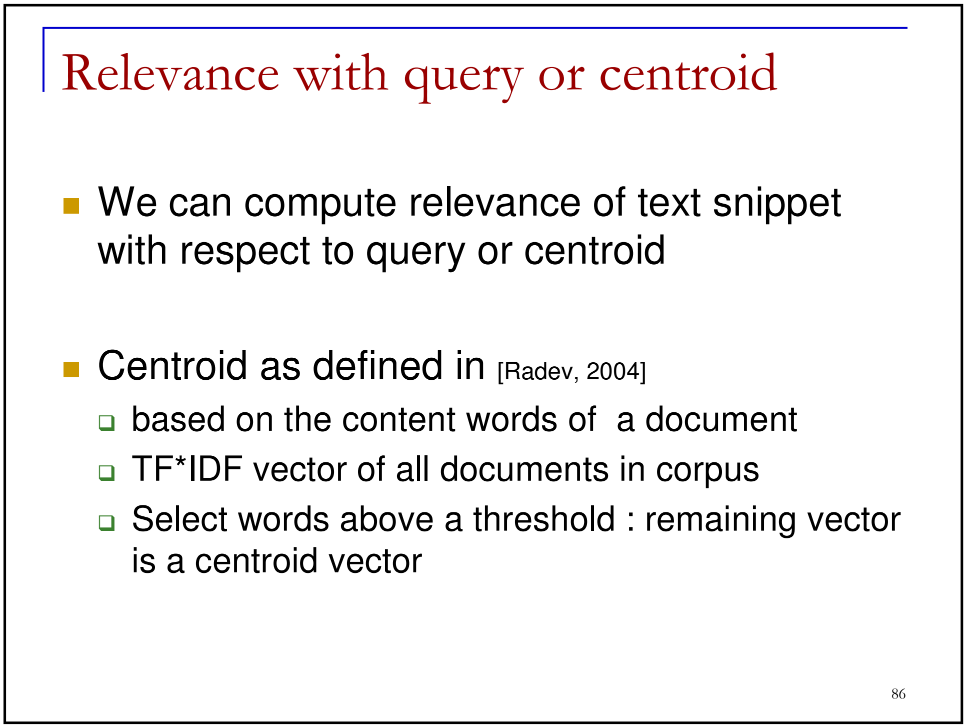


- Summary should maximize relevant information as well as reduce redundancy



Marginal relevance

- "Marginal Relevance" or "Relevant Novelty"
  - Measure relevance and novelty separately
  - Linearly combine these two measures
- High Marginal relevance if
  - Sentence is relevant to story (significant information)
  - Contains minimal similarity to previously selected sentences (new novel information)
- Maximize Marginal Relevance to get summary that has significant non-redundant information



Relevance with query or centroid

- We can compute relevance of text snippet with respect to query or centroid
- Centroid as defined in [Radev, 2004]
  - based on the content words of a document
  - TF*IDF vector of all documents in corpus
  - Select words above a threshold : remaining vector is a centroid vector

Maximal Marginal Relevance (MMR)

[Carbonell et al., 1998]

$$\text{MMR} \approx Argmax_{(D_i \in R-S)}[\lambda(Sim_1(D_i,Q)) - (1-\lambda) max_{(D_j \in S)} Sim_2(D_i,D_j)]$$

- Q – document centroid/user query
- D – document collection
- R – ranked listed
- S – subset of documents in R already selected
- Sim – similarity metric
- Lambda =1 produces most significant ranked list
- Lambda = 0 produces most diverse ranked list

MMR based Summarization [Zechner, 2000]

Iteratively select next sentence

$$\text{Next Sentence} = \begin{array}{l} \arg\max\limits_{t_{nr,j}}(\lambda sim_1(t_{nr,j}, \text{centroid}) \\ -(1-\lambda)\max\limits_{t_{r,k}} sim_2(t_{nr,j}, t_{r,k})) \end{array}$$

Frequency Vector of all content words

$$sim_1 = \vec{\text{tf}}_s \vec{\text{tf}}_t \quad \text{or} \quad \frac{\vec{\text{tf}}_s \vec{\text{tf}}_t}{\left|\vec{\text{tf}}_s\right| \left|\vec{\text{tf}}_t\right|}$$

$$sim_2 = \frac{\vec{\text{tf}}_{t1} \vec{\text{tf}}_{t2}}{\left|\vec{\text{tf}}_{t1}\right| \left|\vec{\text{tf}}_{t2}\right|}$$

$$\text{tf}_{i,s} = 0.5 + 0.5\frac{f_{i,s}}{f_{smax}} \quad \text{or} \quad 1 + \log f_{i,s} \quad \text{or} \quad f_{i,s}$$

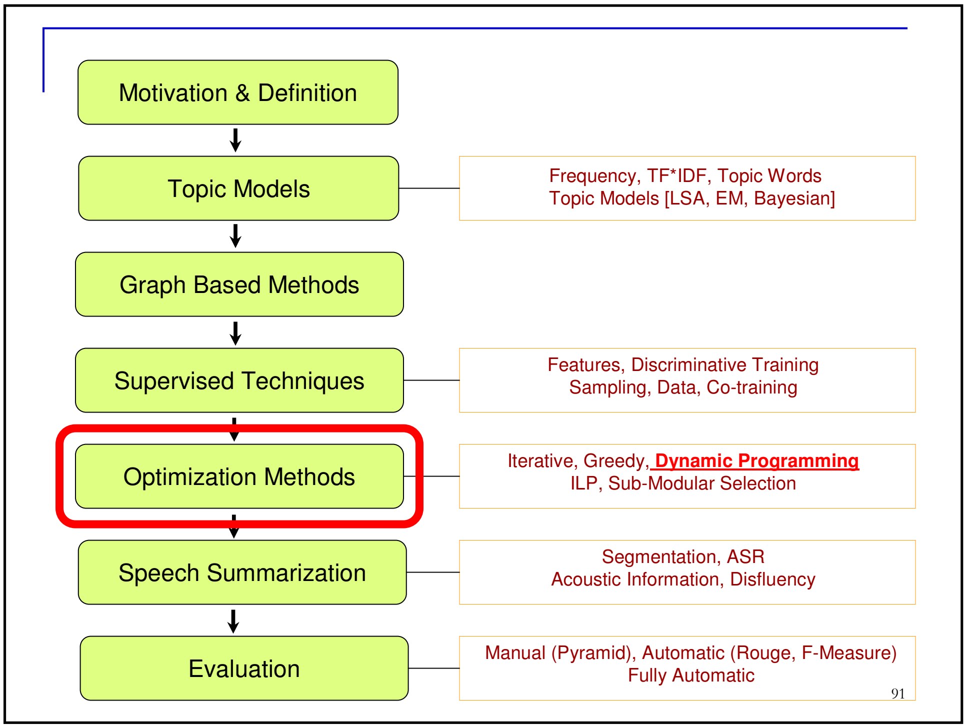
MMR based summarization

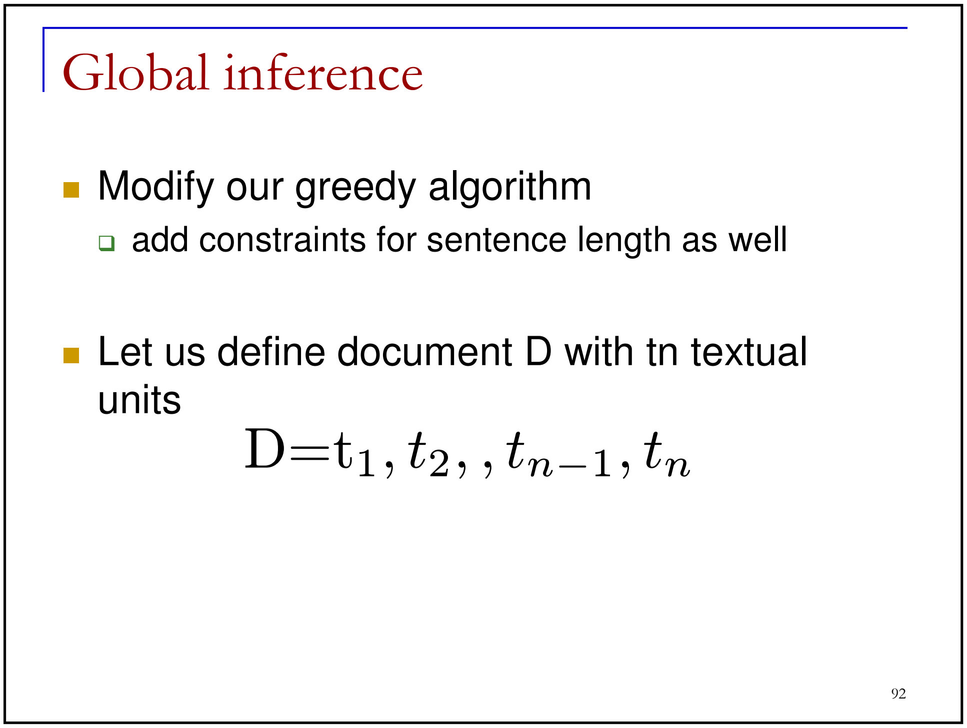
- Why this iterative sentence selection process works?
  - 1st Term: Find relevant sentences similar to centroid of the document
  - 2nd Term: Find redundancy – sentences that are similar to already selected sentences are not selected

Sentence selection in MMR

- MMR is an iterative sentence selection process
  - decision made for each sentence
  - Is this selected sentence globally optimal?

Sentence with same level of relevance but shorter may not be selected if a longer relevant sentence is already selected

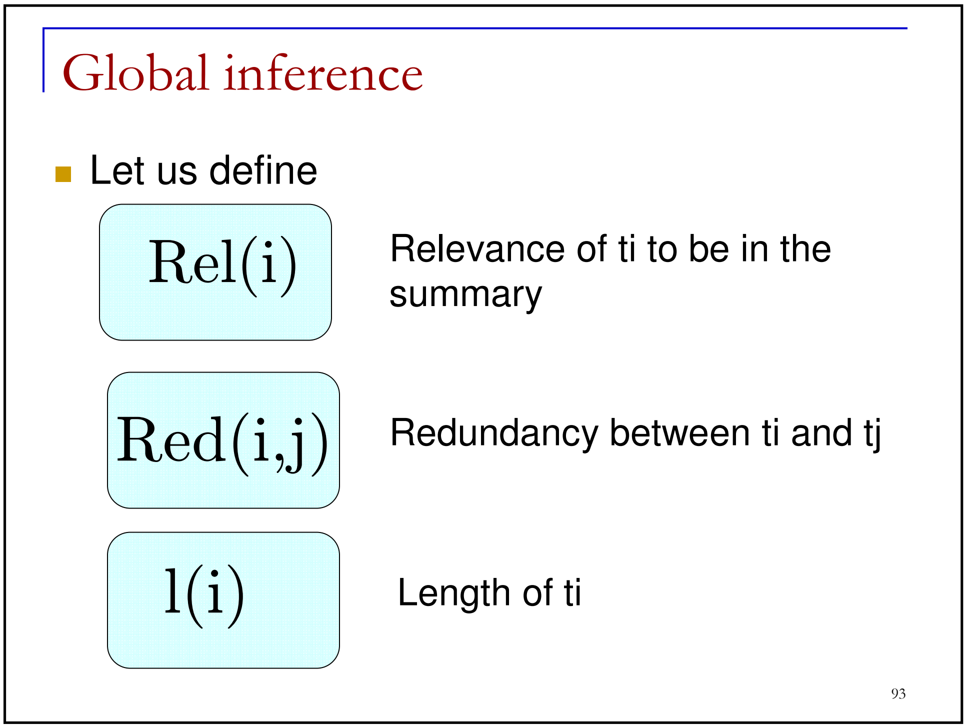

Motivation & Definition
Topic Models
Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]
Graph Based Methods
Supervised Techniques
Features, Discriminative Training
Sampling, Data, Co-training
Optimization Methods
Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection
Speech Summarization
Segmentation, ASR
Acoustic Information, Disfluency
Evaluation
Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic



Global inference

- Modify our greedy algorithm
  - add constraints for sentence length as well
- Let us define document D with tn textual units

$$D=t_1, t_2, , t_{n-1}, t_n$$



Global inference

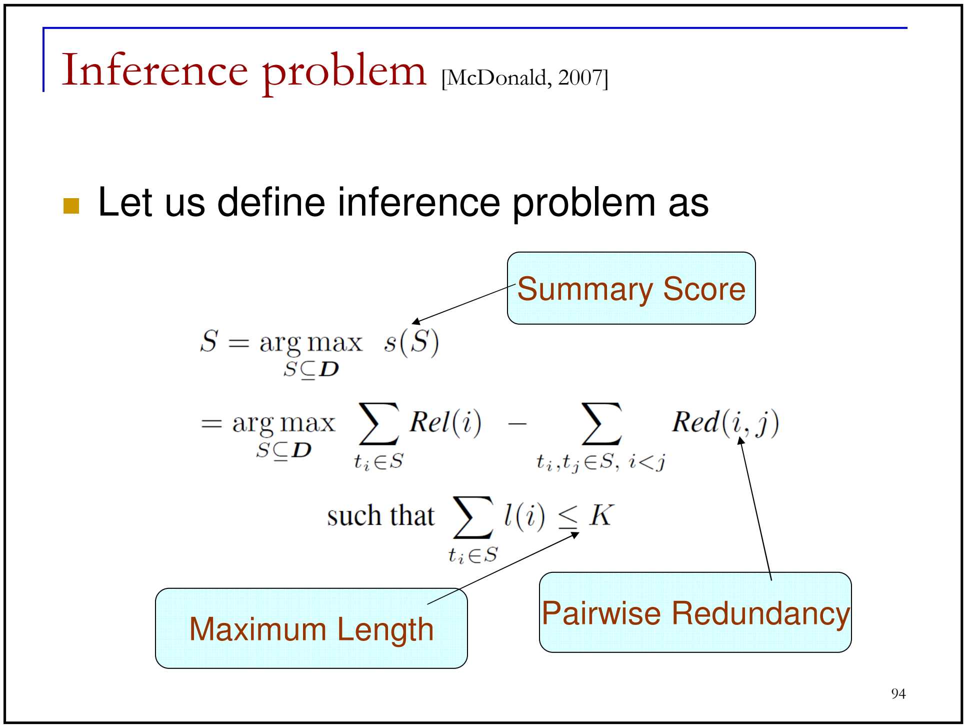
- Let us define

$Rel(i)$ Relevance of ti to be in the summary

$Red(i,j)$ Redundancy between ti and tj

$l(i)$ Length of ti

93



Inference problem [McDonald, 2007]

- Let us define inference problem as

Summary Score

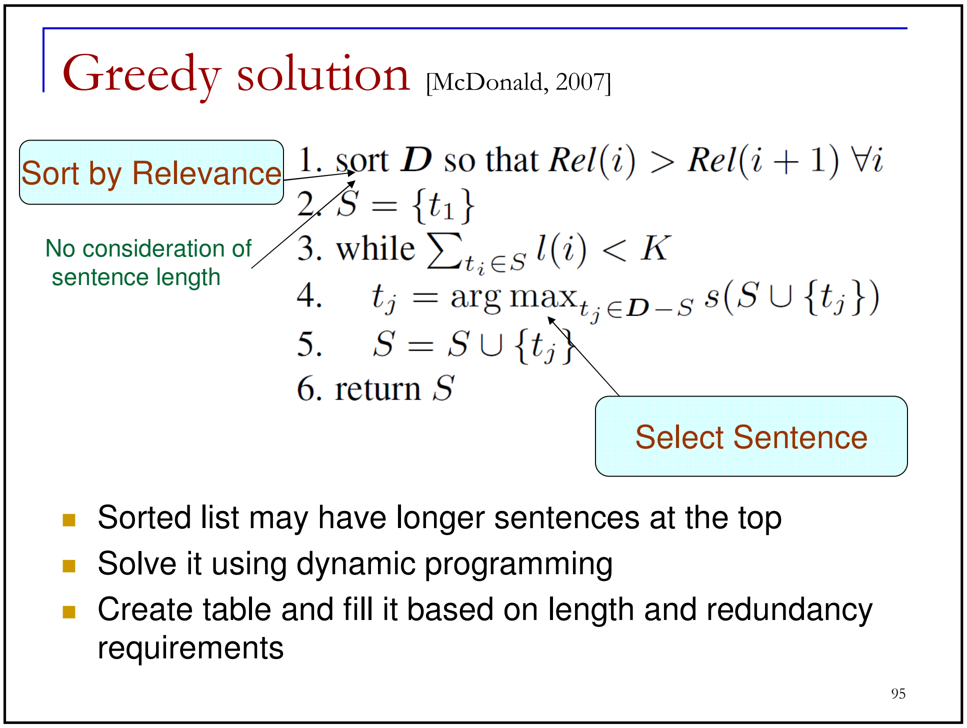
$$
\begin{aligned}
S &= \arg\max_{S\subseteq \boldsymbol{D}} \quad s(S) \\
&= \arg\max_{S\subseteq \boldsymbol{D}} \sum_{t_i\in S} \mathit{Rel}(i) \; - \sum_{t_i,t_j\in S,\; i<j} \mathit{Red}(i,j)
\end{aligned}
$$

$$\text{such that} \sum_{t_i\in S} l(i) \le K$$

Maximum Length

Pairwise Redundancy

94



Greedy solution [McDonald, 2007]

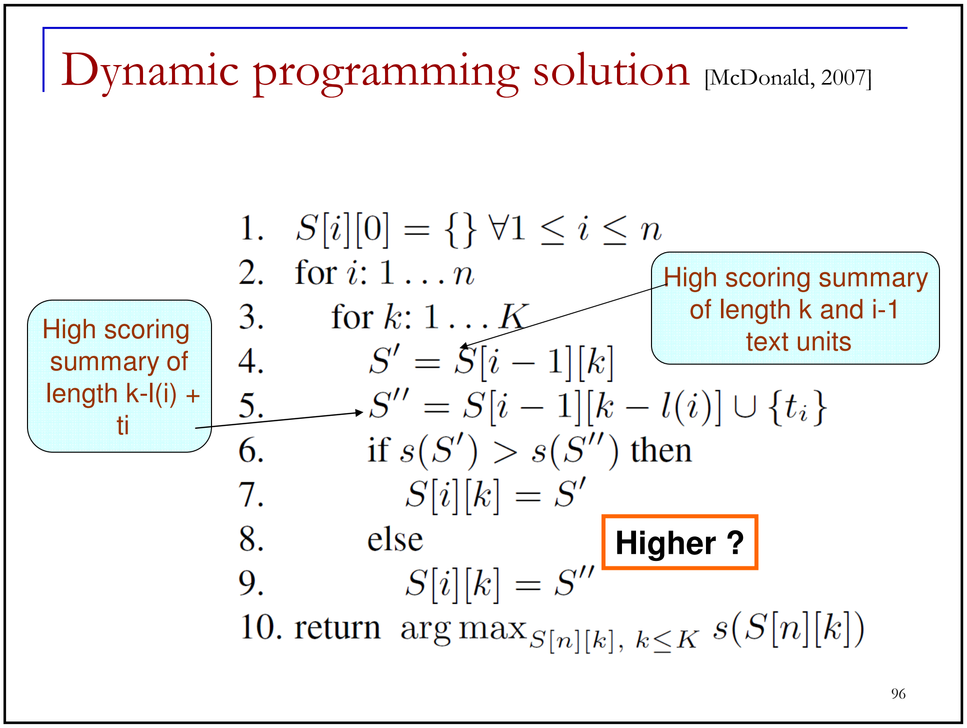
```
sort $\boldsymbol{D}$ so that $Rel(i) > Rel(i+1)\ \forall i$
$S = \{t_1\}$
while $\sum_{t_i \in S} l(i) < K$
    $t_j = \arg\max_{t_j \in \boldsymbol{D}-S} s(S \cup \{t_j\})$
    $S = S \cup \{t_j\}$
return $S$
```

Sort by Relevance

No consideration of sentence length

Select Sentence

- Sorted list may have longer sentences at the top
- Solve it using dynamic programming
- Create table and fill it based on length and redundancy requirements



Dynamic programming solution [McDonald, 2007]

```
$S[i][0] = \{\}\ \forall 1 \le i \le n$
for $i$: $1 \ldots n$
    for $k$: $1 \ldots K$
        $S' = S[i-1][k]$
        $S'' = S[i-1][k-l(i)] \cup \{t_i\}$
        if $s(S') > s(S'')$ then
            $S[i][k] = S'$
        else
            $S[i][k] = S''$
return $\arg\max_{S[n][k],\ k \le K} s(S[n][k])$
```

High scoring summary of length k and i-1 text units

High scoring summary of length k-l(i) + ti

Higher ?

Dynamic programming algorithm [McDonald, 2007]

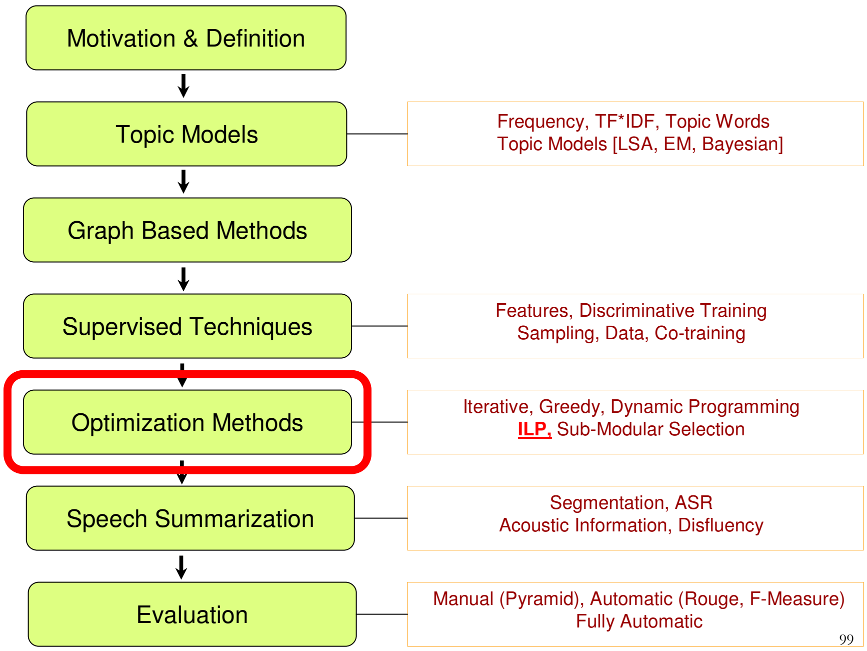
- Better than the previously shown greedy algorithm
- Maximizes the space utilization by not inserting longer sentences
- These are still approximate algorithms: performance loss?

Inference algorithms comparison

[McDonald, 2007]

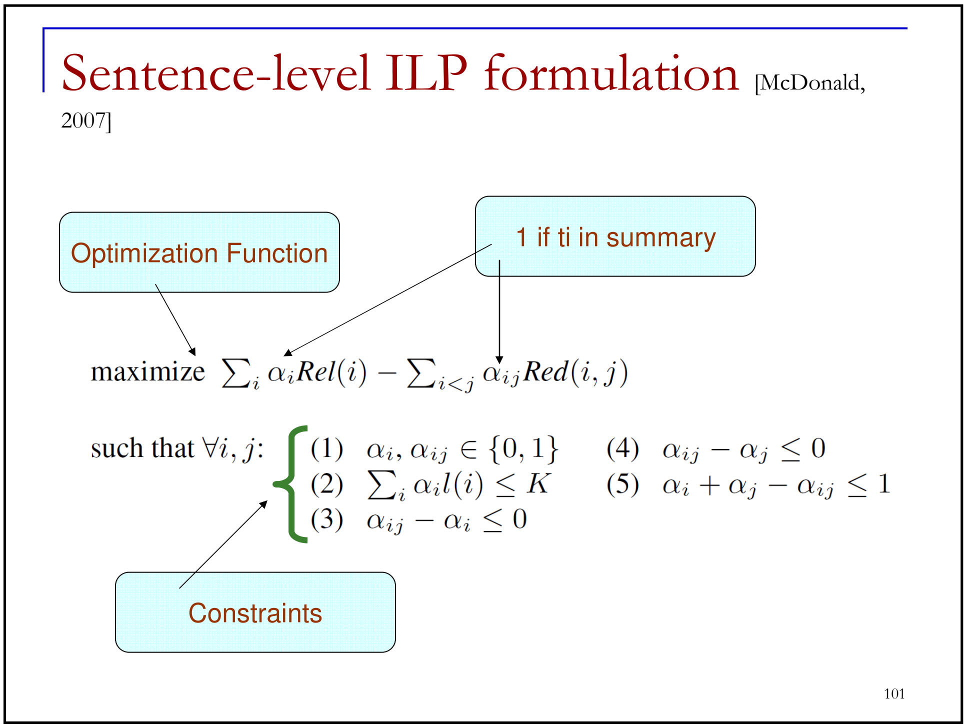
	Sentence Length		
System	50	100	200
Baseline	26.6/5.3	33.0/6.8	39.4/9.6
Greedy	26.8/5.1	33.5/6.9	40.1/9.5
Dynamic Program	27.9/5.9	34.8/7.3	41.2/10.0

Summarization results: Rouge-1/Rouge-2


Motivation & Definition
Topic Models
Graph Based Methods
Supervised Techniques
Optimization Methods
Speech Summarization
Evaluation
Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]
Features, Discriminative Training
Sampling, Data, Co-training
Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection
Segmentation, ASR
Acoustic Information, Disfluency
Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic

Integer Linear Programming (ILP) [Gillick and Favre, 2009; Gillick et al., 2009; McDonald, 2007]

- Greedy algorithm is an approximate solution
- Use exact solution algorithm with ILP (scaling issues though)
- ILP is constrained optimization problem
  - Cost and constraints are linear in a set of integer variables
- Many solvers on the web
- Define the constraints based on relevance and redundancy for summarization
  - Sentence based ILP
  - N-gram based ILP



Sentence-level ILP formulation [McDonald, 2007]

Optimization Function

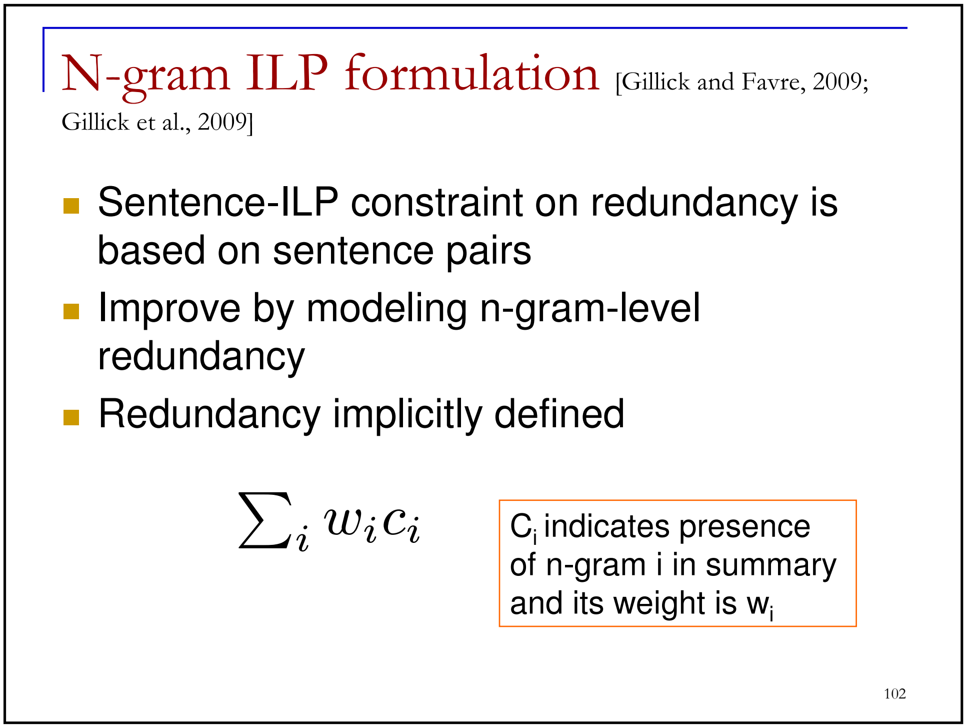
1 if ti in summary

$$\text{maximize} \ \sum_i \alpha_i Rel(i) - \sum_{i<j} \alpha_{ij} Red(i,j)$$

$$\text{such that } \forall i,j: \left\{ \begin{array}{llll} (1) & \alpha_i, \alpha_{ij} \in \{0,1\} & (4) & \alpha_{ij} - \alpha_j \leq 0 \\ (2) & \sum_i \alpha_i l(i) \leq K & (5) & \alpha_i + \alpha_j - \alpha_{ij} \leq 1 \\ (3) & \alpha_{ij} - \alpha_i \leq 0 & & \end{array} \right.$$

Constraints

101



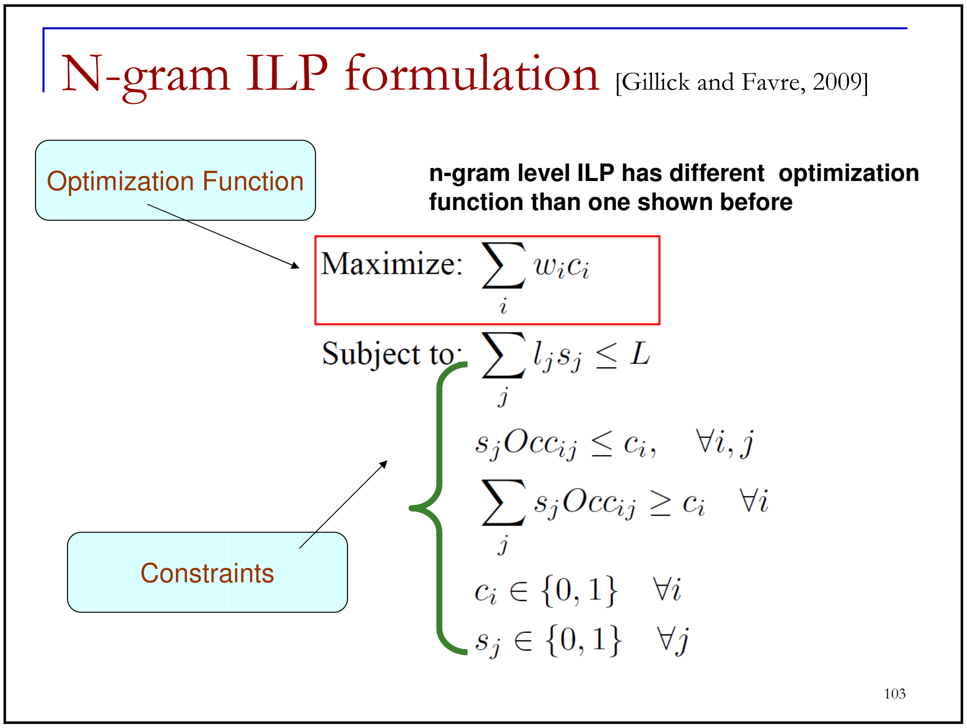
N-gram ILP formulation [Gillick and Favre, 2009; Gillick et al., 2009]

- Sentence-ILP constraint on redundancy is based on sentence pairs
- Improve by modeling n-gram-level redundancy
- Redundancy implicitly defined

$$\sum_i w_i c_i$$

C_i indicates presence of n-gram i in summary and its weight is w_i

102



N-gram ILP formulation [Gillick and Favre, 2009]

Optimization Function

n-gram level ILP has different optimization function than one shown before

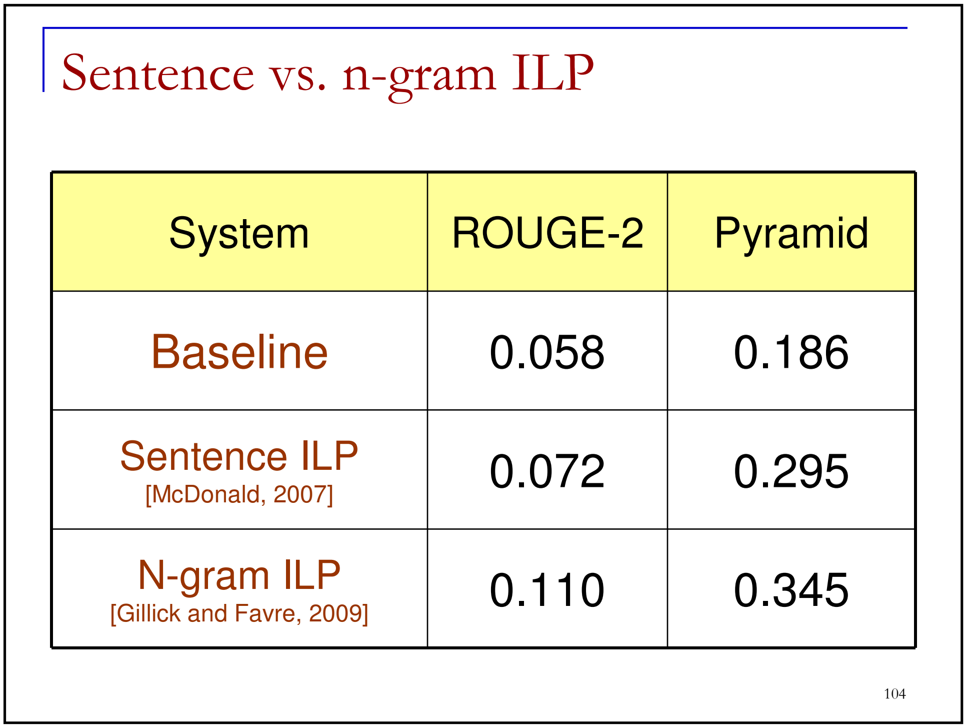
$$\text{Maximize: } \sum_i w_i c_i$$

$$\text{Subject to: } \sum_j l_j s_j \leq L$$

Constraints

$$\left\{\begin{array}{l} s_j Occ_{ij} \leq c_i, \quad \forall i, j \\ \sum_j s_j Occ_{ij} \geq c_i \quad \forall i \\ c_i \in \{0, 1\} \quad \forall i \\ s_j \in \{0, 1\} \quad \forall j \end{array}\right.$$

103



Sentence vs. n-gram ILP

System	ROUGE-2	Pyramid
Baseline	0.058	0.186
Sentence ILP [McDonald, 2007]	0.072	0.295
N-gram ILP [Gillick and Favre, 2009]	0.110	0.345

104

Other optimization based summarization algorithms

- Submodular selection [Lin et al., 2009]
  - Submodular set functions for optimization
- Modified greedy algorithm [Filatova, 2004]
  - Event based features
- Stack decoding algorithm [Yih et al., 2007]
  - Multiple stacks, each stack represents hypothesis of different length
- A* Search [Aker et al., 2010]
  - Use scoring and heuristic functions

Submodular selection for summarization

[Lin et al., 2009]

- Summarization Setup
  - V – set of all sentences in document
  - S – set of extraction sentences
  - f(.) scores the quality of the summary
- Submodularity been used in solving many optimization problems in near polynomial time
- For summarization:

Select subset S (sentences) representative of V given the constraint |S| =< K (budget)

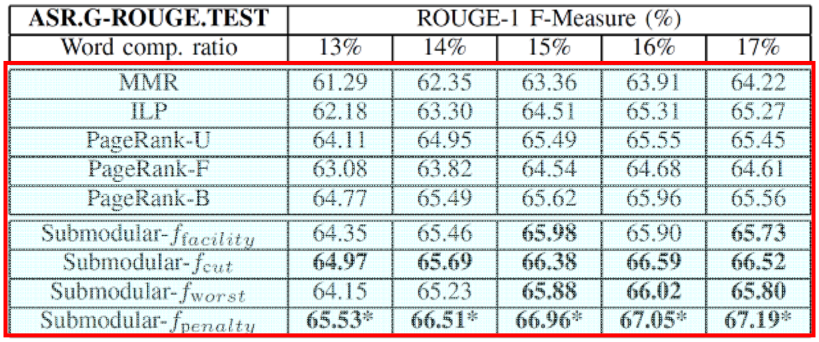
Submodular selection [Lin et al., 2009]

- If V are nodes in a Graph G=(V,E) representing sentences
- And E represents edges (i,j) such that w(i,j) represents similarity between sentences i and j
- Introduce submodular set functions which measures “representative” S of entire set V
- [Lin et al., 2009] presented 4 submodular set functions

Submodular selection for summarization

[Lin et al., 2009]

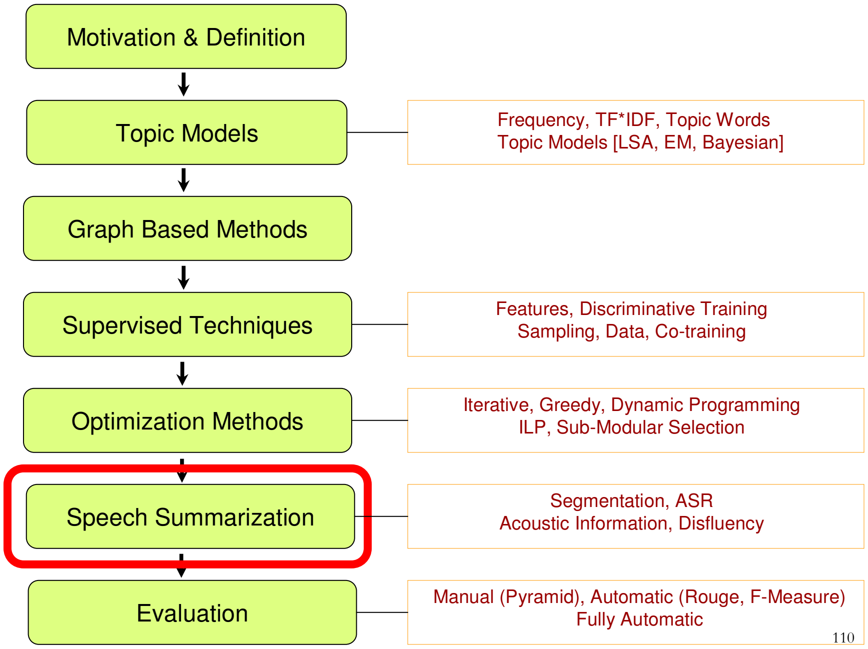
Comparison of results using different methods



ASR.G-ROUGE.TEST	ROUGE-1 F-Measure (%)				
Word comp. ratio	13%	14%	15%	16%	17%
MMR	61.29	62.35	63.36	63.91	64.22
ILP	62.18	63.30	64.51	65.31	65.27
PageRank-U	64.11	64.95	65.49	65.55	65.45
PageRank-F	63.08	63.82	64.54	64.68	64.61
PageRank-B	64.77	65.49	65.62	65.96	65.56
Submodular-$f_{facility}$	64.35	65.46	**65.98**	65.90	**65.73**
Submodular-f_{cut}	**64.97**	**65.69**	**66.38**	**66.59**	**66.52**
Submodular-f_{worst}	64.15	65.23	**65.88**	**66.02**	**65.80**
Submodular-$f_{penalty}$	**65.53***	**66.51***	**66.96***	**67.05***	**67.19***

Review: optimization methods

- Global optimization methods have shown to be superior than 2-step selection process and reduce redundancy
- 3 parameters are optimized together
  - Relevance
  - Redundancy
  - Length
- Various Algorithms for Global Inference
  - Greedy
  - Dynamic Programming
  - Integer Linear Programming
  - Submodular Selection



Motivation & Definition

Topic Models — Frequency, TF*IDF, Topic Words; Topic Models [LSA, EM, Bayesian]

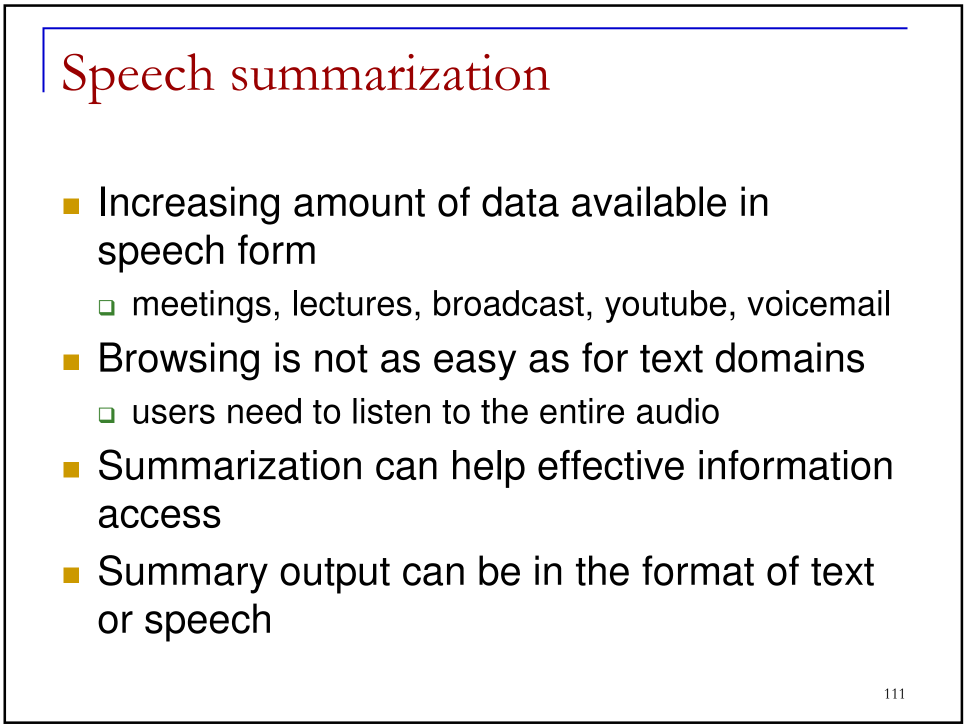
Graph Based Methods

Supervised Techniques — Features, Discriminative Training; Sampling, Data, Co-training

Optimization Methods — Iterative, Greedy, Dynamic Programming; ILP, Sub-Modular Selection

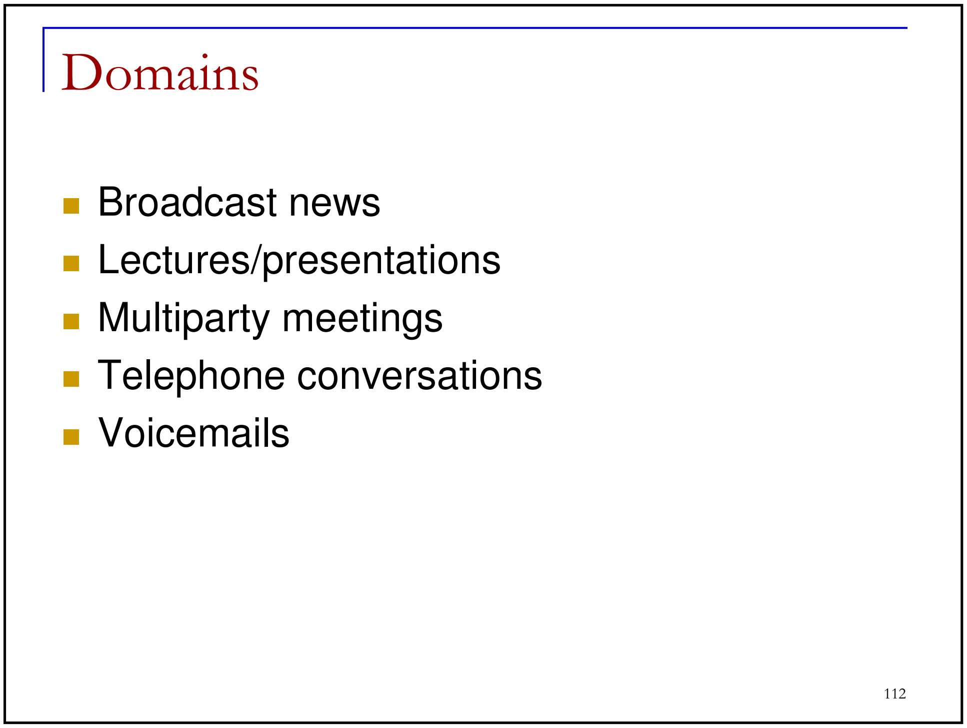
Speech Summarization — Segmentation, ASR; Acoustic Information, Disfluency

Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure); Fully Automatic



Speech summarization

- Increasing amount of data available in speech form
  - meetings, lectures, broadcast, youtube, voicemail
- Browsing is not as easy as for text domains
  - users need to listen to the entire audio
- Summarization can help effective information access
- Summary output can be in the format of text or speech



Domains

- Broadcast news
- Lectures/presentations
- Multiparty meetings
- Telephone conversations
- Voicemails

Example

Meeting transcripts and summary sentences (in red)

me010	**there there are a variety of ways of doing it**
me010	uh let me just mention something that i don't want to pursue today
me010	which is there are technical ways of doing it
me010	uh i- i slipped a paper to bhaskara and about noisy-or's and noisy-maxes
me010	and
me010	there're ways to uh sort of back off on the purity of your bayes-net-edness
me003	mmm
me010	uh so if you co- you could ima- and i-
me010	now I don't know that any of these actually apply in this case
me010	but there is some technology you could try to apply
me003	**so it's possible that we could do something like a summary node of some sort that**
me010	yeah

Broadcast news transcripts and summary (in red)

california's strained power grid is getting a boost today which might help increasingly taxed power supplies

a unit at diablo canyon nuclear plant is expected to resume production today

it had been shut down for maintenance

coupled with another unit, it can provide enough power for about 2 million people

meanwhile, a cold snap in the pacific northwest is putting an added strain on power supplies

the area shares power across many states

energy officials are offering tips to conserve electricity, they say, to delay holiday lighting until after at night

set your thermostat at 68 degrees when you're home, 55 degrees when you're away

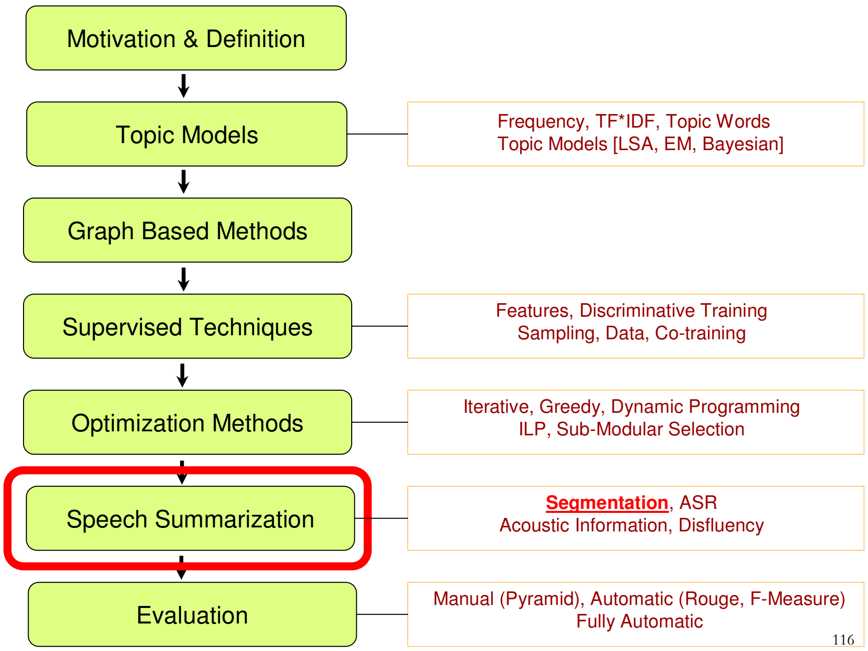
try to use electrical appliances before p.m. and after p.m. and turn off computers, copiers and lights when they're not being used

Speech vs. text summarization: similarities

- When high quality transcripts are available
  - Not much different from text summarization
  - Many similar approaches have been used
  - Some also incorporate acoustic information
- For genres like broadcast news, style is also similar to text domains

Speech vs. text summarization: differences

- Challenges in speech summarization
  - Speech recognition errors can be very high
  - Sentences are not as well formed as in most text domains: disfluencies, ungrammatical
  - There are not clearly defined sentences
  - Information density is also low (off-topic discussions, chit chat, etc.)
  - Multiple participants



Motivation & Definition

Topic Models — Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]

Graph Based Methods

Supervised Techniques — Features, Discriminative Training
Sampling, Data, Co-training

Optimization Methods — Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection

Speech Summarization — **Segmentation**, ASR
Acoustic Information, Disfluency

Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic

What should be extraction units in speech summarization?

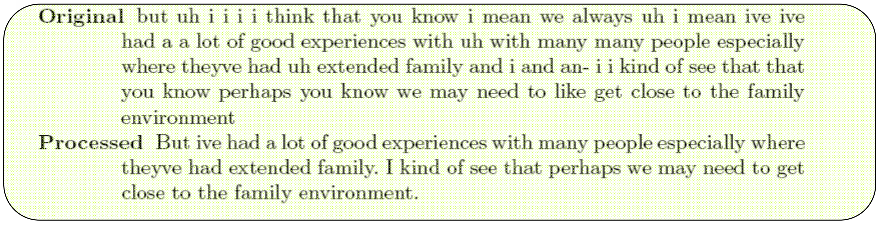
- Text domain
  - Typically use sentences (based on punctuation marks)
- Speech domain
  - Sentence information is not available
  - Sentences are not as clearly defined

Utterance from previous example:

there there are a variety of ways of doing it uh let me just mention something that i don't want to pursue today which is there are technical ways of doing it

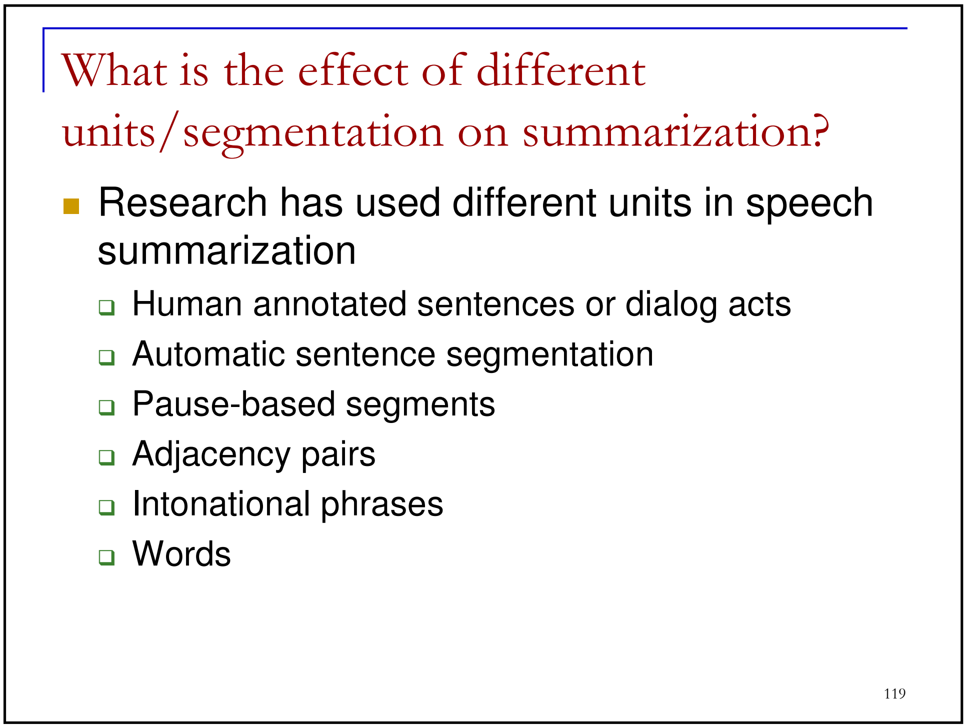
Automatic sentence segmentation (side note)

- For a word boundary, determine whether it's a sentence boundary
- Different approaches:
  - Generative: HMM
  - Discriminative: SVM, boosting, maxent, CRF
  - Information used: word n-gram, part-of-speech, parsing information, acoustic info (pause, pitch, energy)



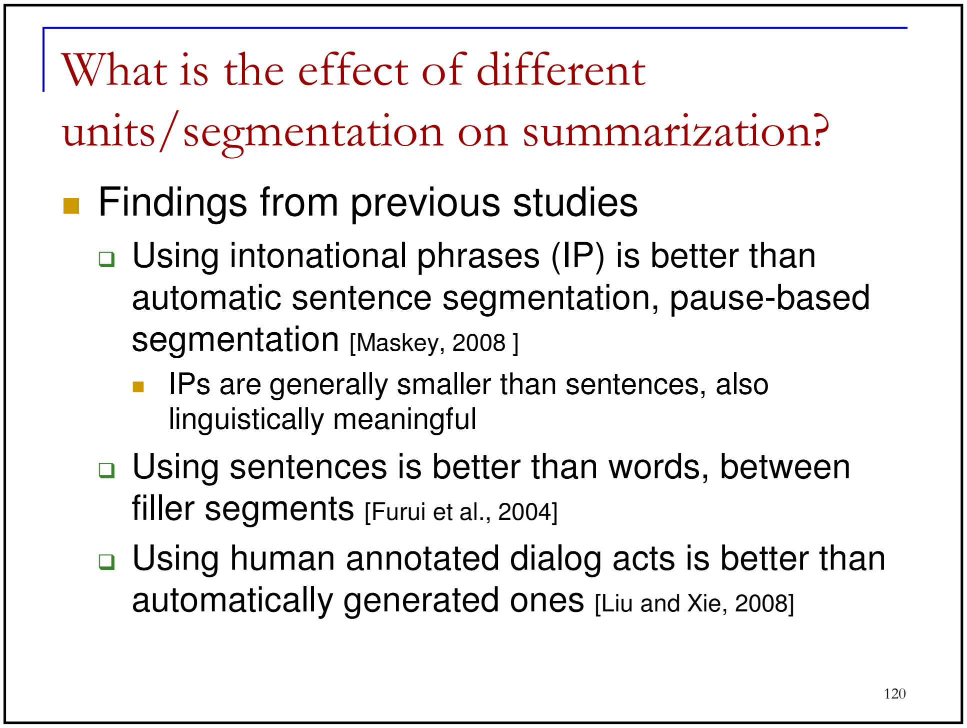
Original but uh i i i i think that you know i mean we always uh i mean ive ive had a a lot of good experiences with uh with many many people especially where theyve had uh extended family and i and an- i i kind of see that that you know perhaps you know we may need to like get close to the family environment

Processed But ive had a lot of good experiences with many people especially where theyve had extended family. I kind of see that perhaps we may need to get close to the family environment.



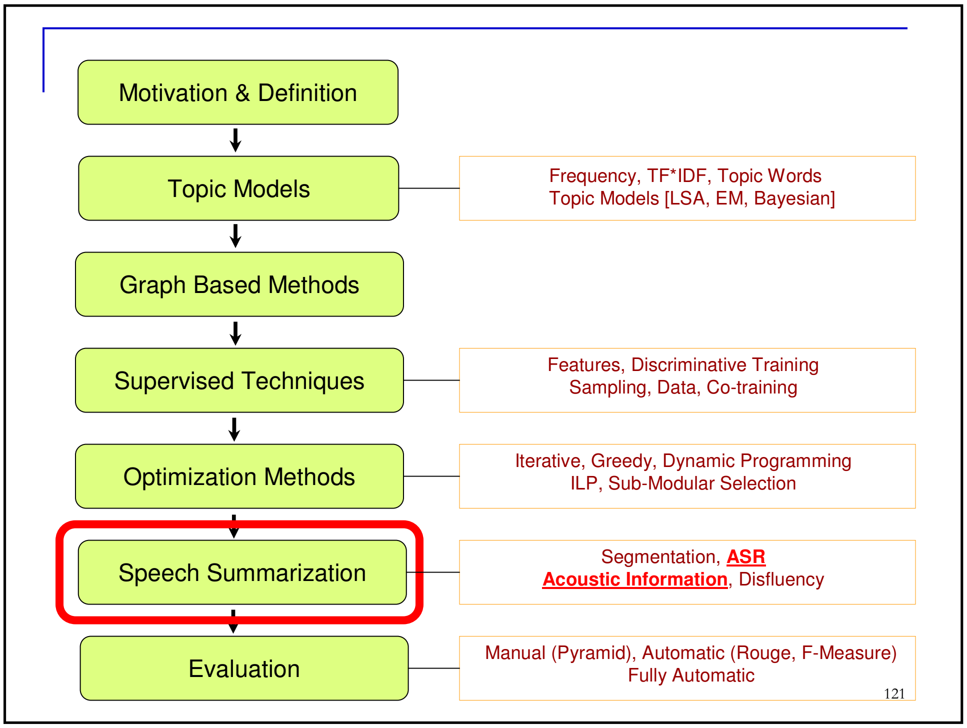
What is the effect of different units/segmentation on summarization?

- Research has used different units in speech summarization
  - Human annotated sentences or dialog acts
  - Automatic sentence segmentation
  - Pause-based segments
  - Adjacency pairs
  - Intonational phrases
  - Words



What is the effect of different units/segmentation on summarization?

- Findings from previous studies
  - Using intonational phrases (IP) is better than automatic sentence segmentation, pause-based segmentation [Maskey, 2008]
    - IPs are generally smaller than sentences, also linguistically meaningful
  - Using sentences is better than words, between filler segments [Furui et al., 2004]
  - Using human annotated dialog acts is better than automatically generated ones [Liu and Xie, 2008]



Motivation & Definition

Topic Models — Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]

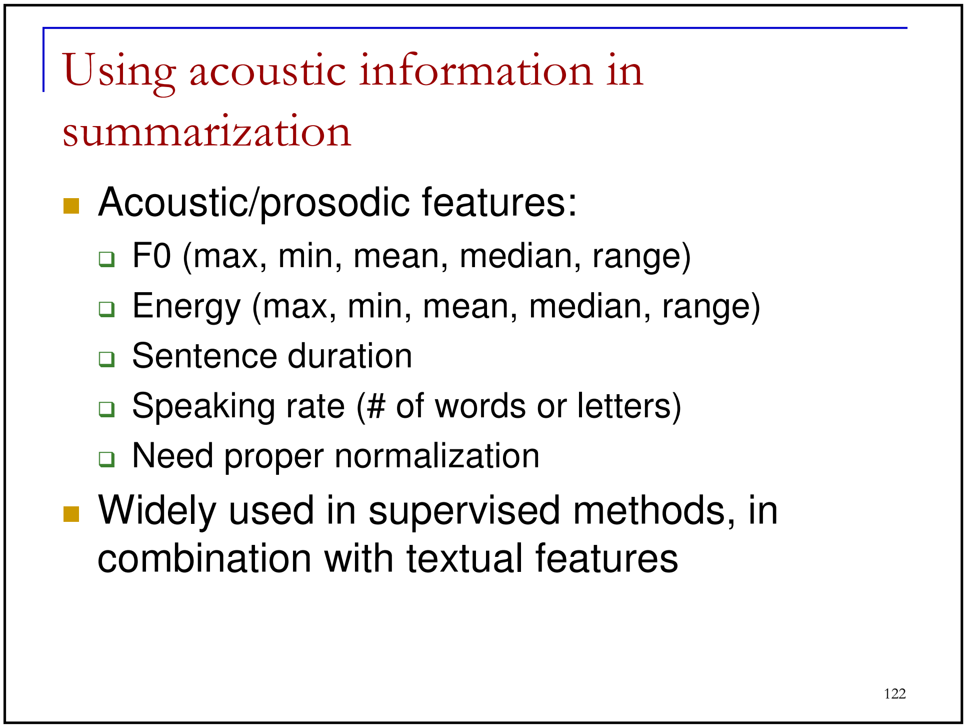
Graph Based Methods

Supervised Techniques — Features, Discriminative Training
Sampling, Data, Co-training

Optimization Methods — Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection

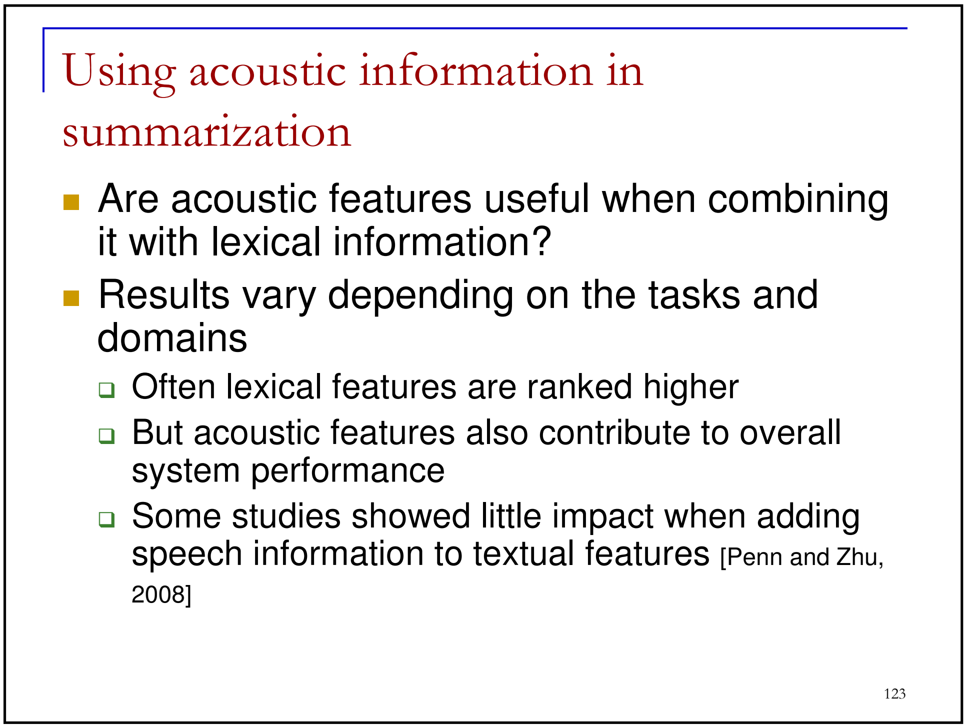
Speech Summarization — Segmentation, **ASR**
Acoustic Information, Disfluency

Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic



Using acoustic information in summarization

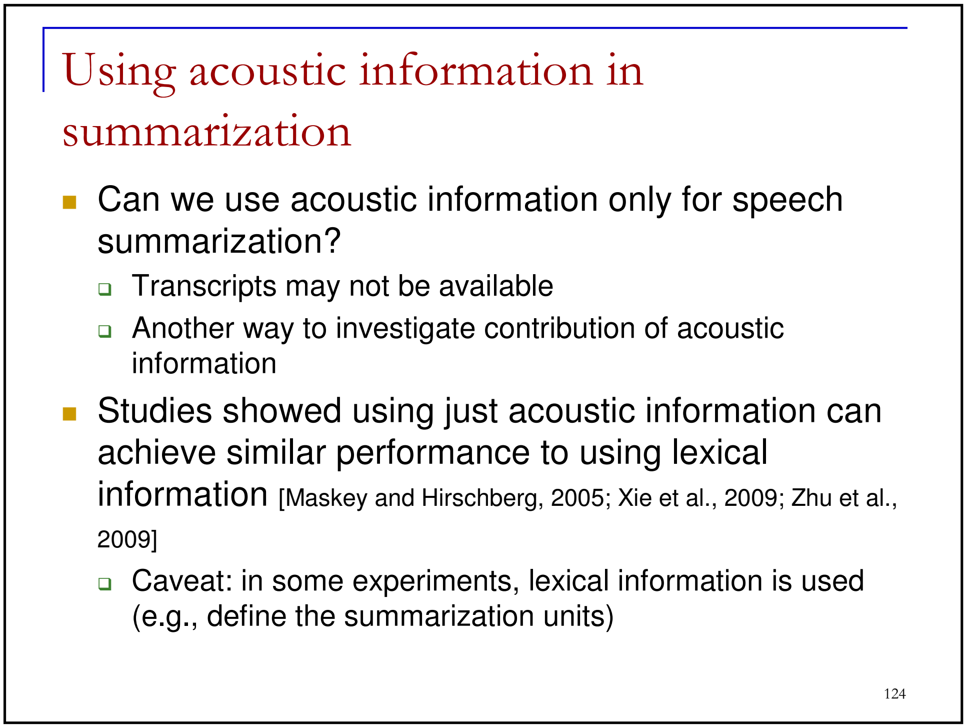
- Acoustic/prosodic features:
  - F0 (max, min, mean, median, range)
  - Energy (max, min, mean, median, range)
  - Sentence duration
  - Speaking rate (# of words or letters)
  - Need proper normalization
- Widely used in supervised methods, in combination with textual features



Using acoustic information in summarization

- Are acoustic features useful when combining it with lexical information?
- Results vary depending on the tasks and domains
  - Often lexical features are ranked higher
  - But acoustic features also contribute to overall system performance
  - Some studies showed little impact when adding speech information to textual features [Penn and Zhu, 2008]

123



Using acoustic information in summarization

- Can we use acoustic information only for speech summarization?
  - Transcripts may not be available
  - Another way to investigate contribution of acoustic information
- Studies showed using just acoustic information can achieve similar performance to using lexical information [Maskey and Hirschberg, 2005; Xie et al., 2009; Zhu et al., 2009]
  - Caveat: in some experiments, lexical information is used (e.g., define the summarization units)

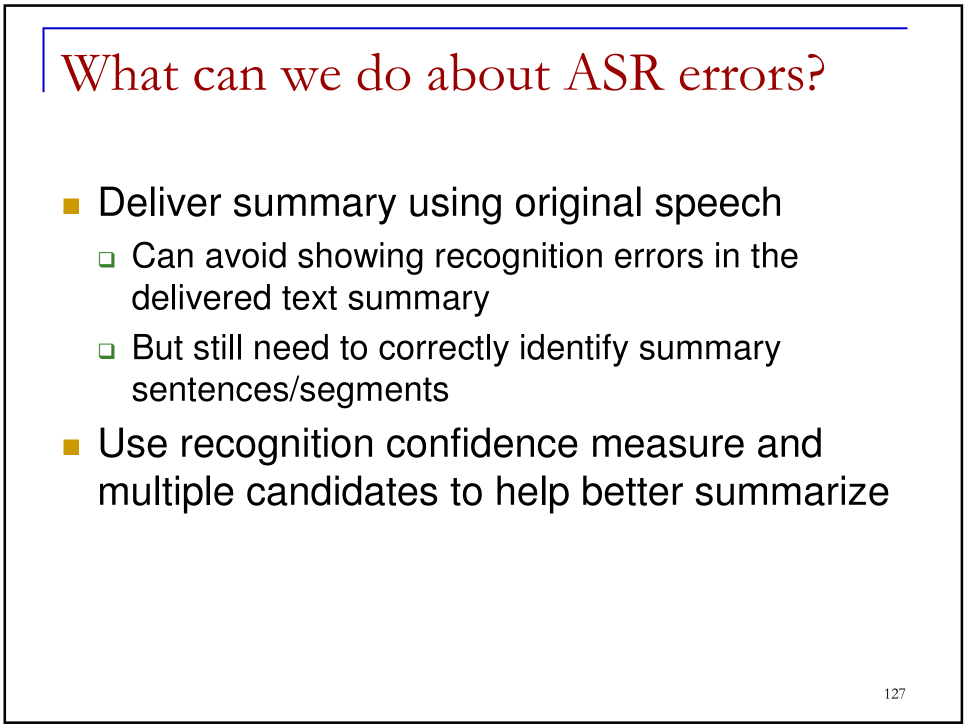
124

Speech recognition errors

- ASR is not perfect, often high word error rate
  - 10-20% for read speech
  - 40% or even higher for conversational speech
- Recognition errors generally have negative impact on summarization performance
  - Important topic indicative words are incorrectly recognized
  - Can affect term weighting and sentence scores

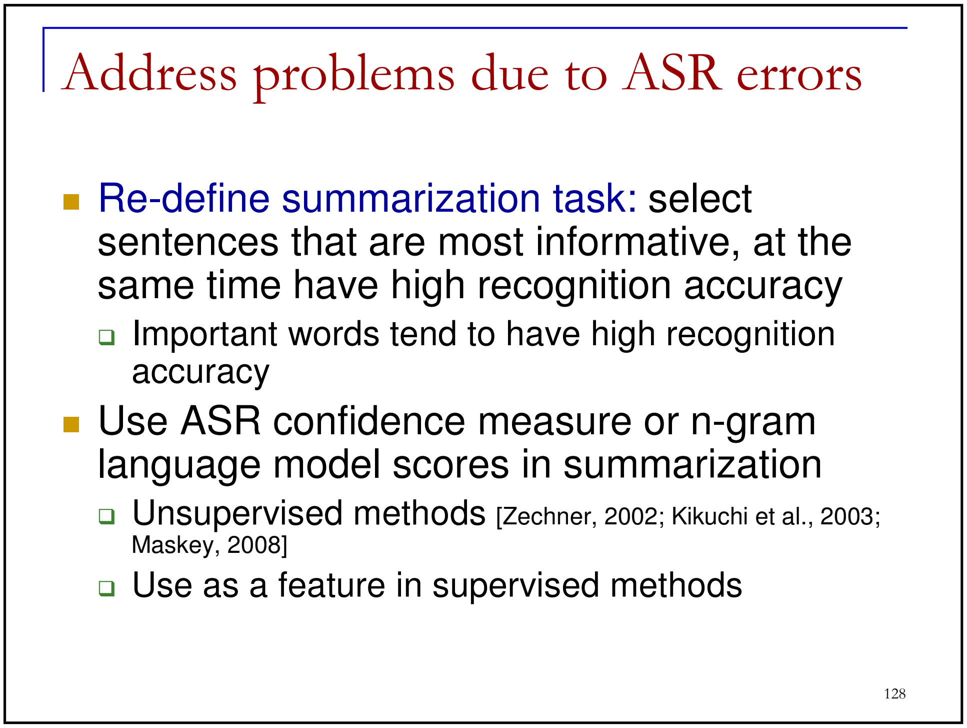
Speech recognition errors

- Some studies evaluated effect of recognition errors on summarization by varying word error rate [Christensen et al., 2003; Penn and Zhu, 2008; Lin et al., 2009]
- Degradation is not much when word error rate is not too low (similar to spoken document retrieval)
  - Reason: better recognition accuracy in summary sentences than overall



What can we do about ASR errors?

- Deliver summary using original speech
  - Can avoid showing recognition errors in the delivered text summary
  - But still need to correctly identify summary sentences/segments
- Use recognition confidence measure and multiple candidates to help better summarize

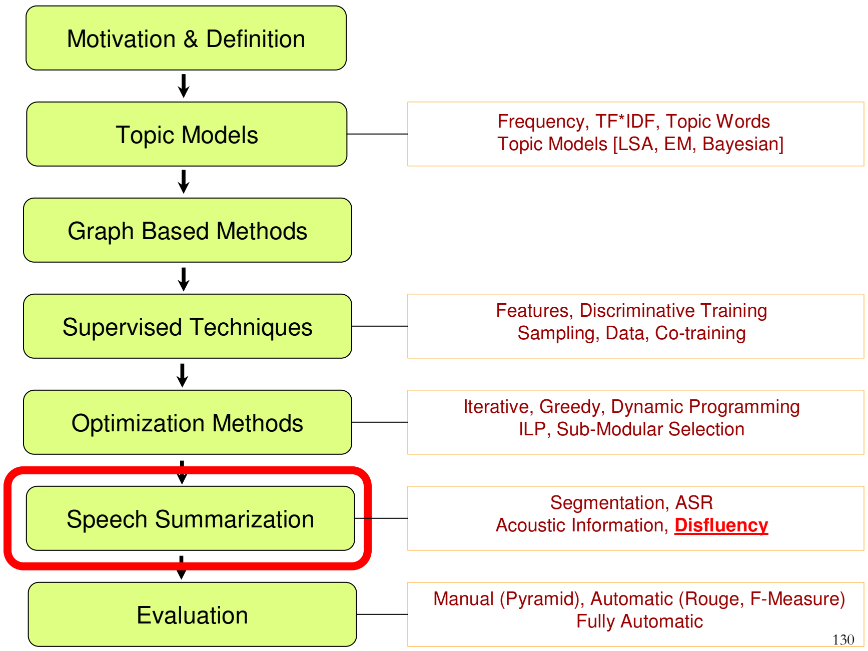


Address problems due to ASR errors

- Re-define summarization task: select sentences that are most informative, at the same time have high recognition accuracy
  - Important words tend to have high recognition accuracy
- Use ASR confidence measure or n-gram language model scores in summarization
  - Unsupervised methods [Zechner, 2002; Kikuchi et al., 2003; Maskey, 2008]
  - Use as a feature in supervised methods

Address problems due to ASR errors

- Use multiple recognition candidates
  - n-best lists [Liu et al., 2010]
  - Lattices [Lin et al., 2010]
  - Confusion network [Xie and Liu, 2010]
    - Use in MMR framework
    - Summarization segment/unit contains all the word candidates (or pruned ones based on probabilities)
    - Term weights (TF, IDF) use candidate's posteriors
    - Improved performance over using 1-best recognition output



Motivation & Definition

Topic Models

Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]

Graph Based Methods

Supervised Techniques

Features, Discriminative Training
Sampling, Data, Co-training

Optimization Methods

Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection

Speech Summarization

Segmentation, ASR
Acoustic Information, **Disfluency**

Evaluation

Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic

Disfluencies and summarization

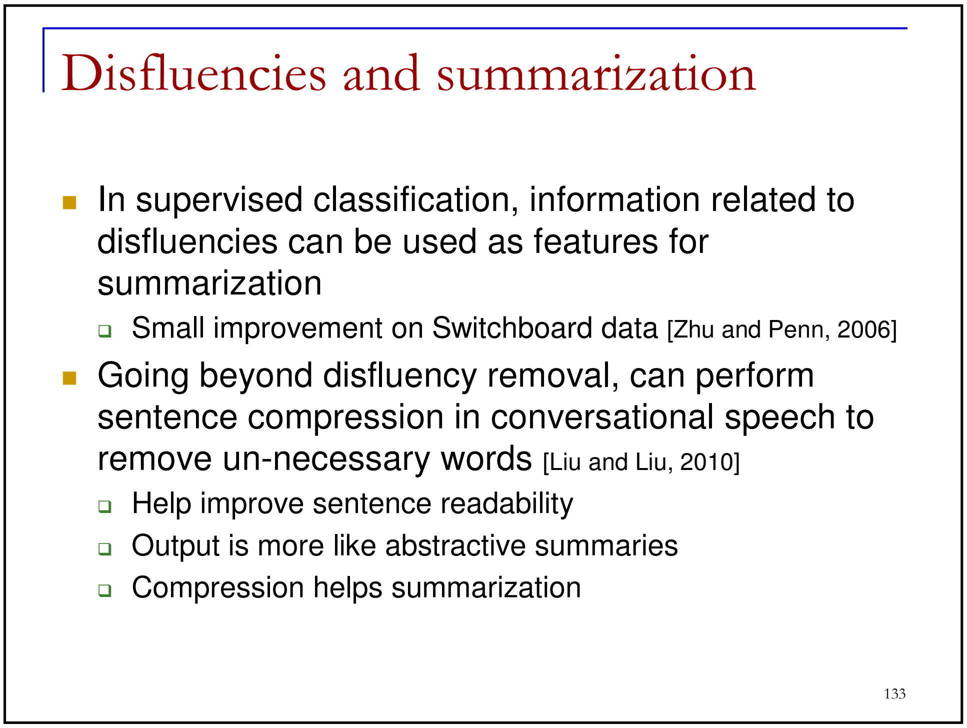
- Disfluencies (filler words, repetitions, revisions, restart, etc) are frequent in conversational speech
  - Example from meeting transcript:

> so so does i- just remind me of what what you were going to do with the what what what what's
>
> y- you just described what you've been doing

- Existence may hurt summarization systems, also affect human readability of the summaries

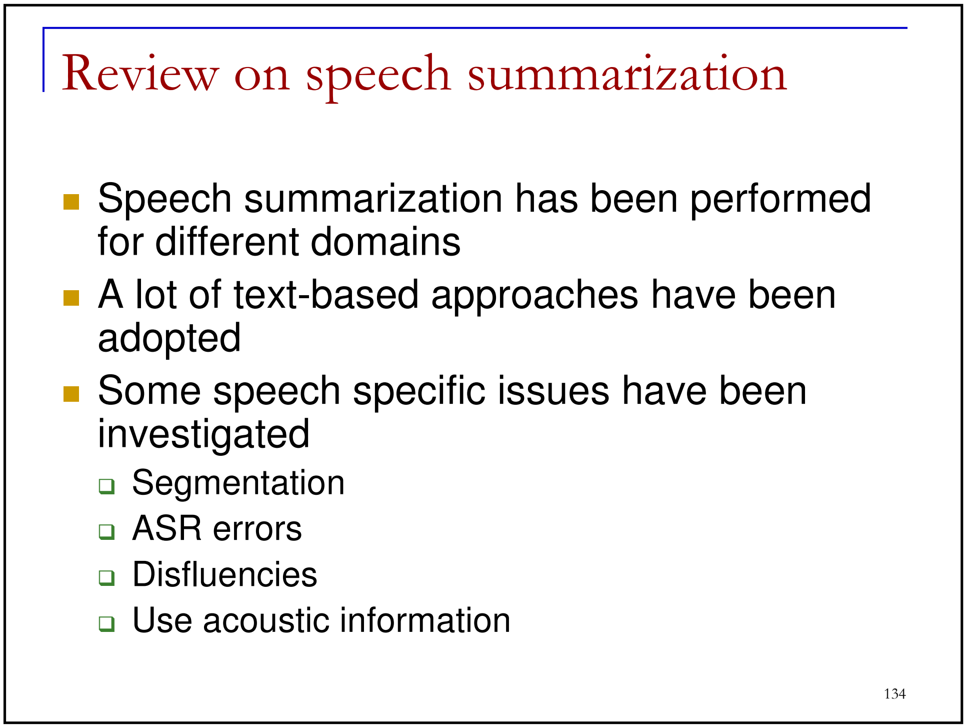
Disfluencies and summarization

- Natural thought: remove disfluenices
- Word-based selection can avoid disfluent words
  - Using n-gram scores tends to select fluent parts [Hori and Furui, 2001]
- Remove disfluencies first, then perform summarization
  - Does it work? not consistent results
    - Small improvement [Maskey, 2008; Zechner, 2002]
    - No improvement [Liu et al., 2007]



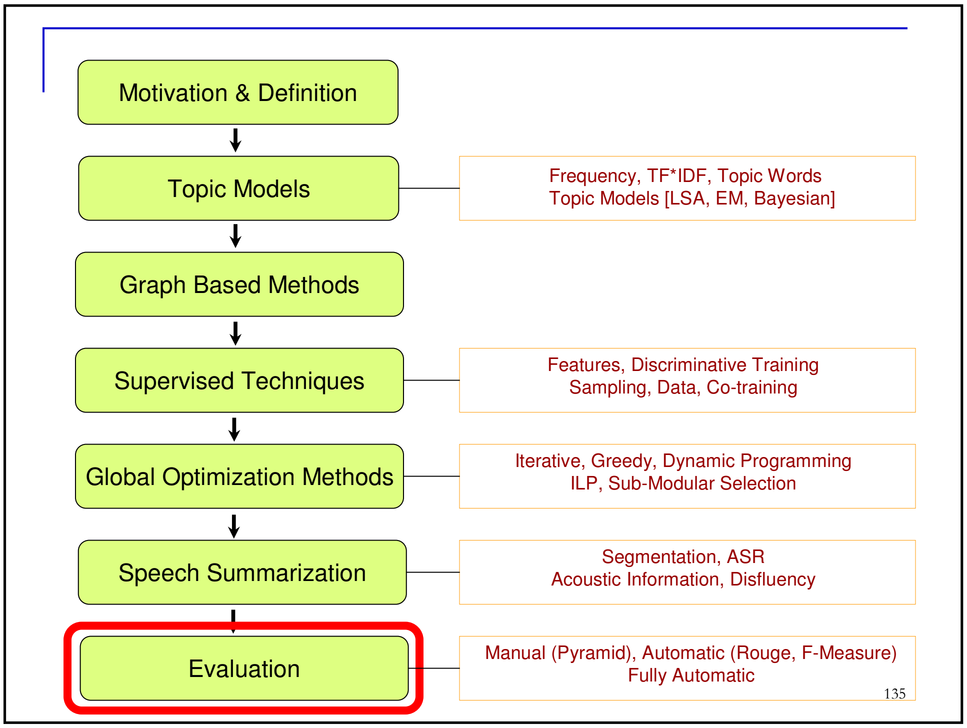
Disfluencies and summarization

- In supervised classification, information related to disfluencies can be used as features for summarization
  - Small improvement on Switchboard data [Zhu and Penn, 2006]
- Going beyond disfluency removal, can perform sentence compression in conversational speech to remove un-necessary words [Liu and Liu, 2010]
  - Help improve sentence readability
  - Output is more like abstractive summaries
  - Compression helps summarization



Review on speech summarization

- Speech summarization has been performed for different domains
- A lot of text-based approaches have been adopted
- Some speech specific issues have been investigated
  - Segmentation
  - ASR errors
  - Disfluencies
  - Use acoustic information



Motivation & Definition

Topic Models — Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]

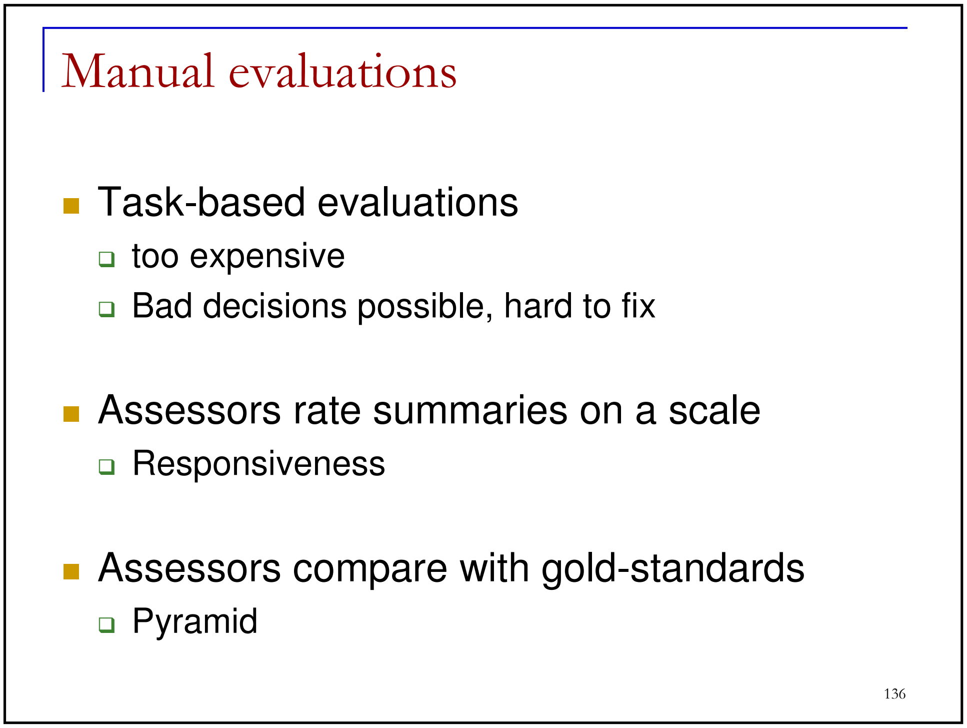
Graph Based Methods

Supervised Techniques — Features, Discriminative Training
Sampling, Data, Co-training

Global Optimization Methods — Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection

Speech Summarization — Segmentation, ASR
Acoustic Information, Disfluency

Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic



Manual evaluations

- Task-based evaluations
  - too expensive
  - Bad decisions possible, hard to fix
- Assessors rate summaries on a scale
  - Responsiveness
- Assessors compare with gold-standards
  - Pyramid

Automatic and fully automatic evaluation

- Automatically compare with gold-standard
  - Precision/recall (sentence level)
  - ROUGE (word level)

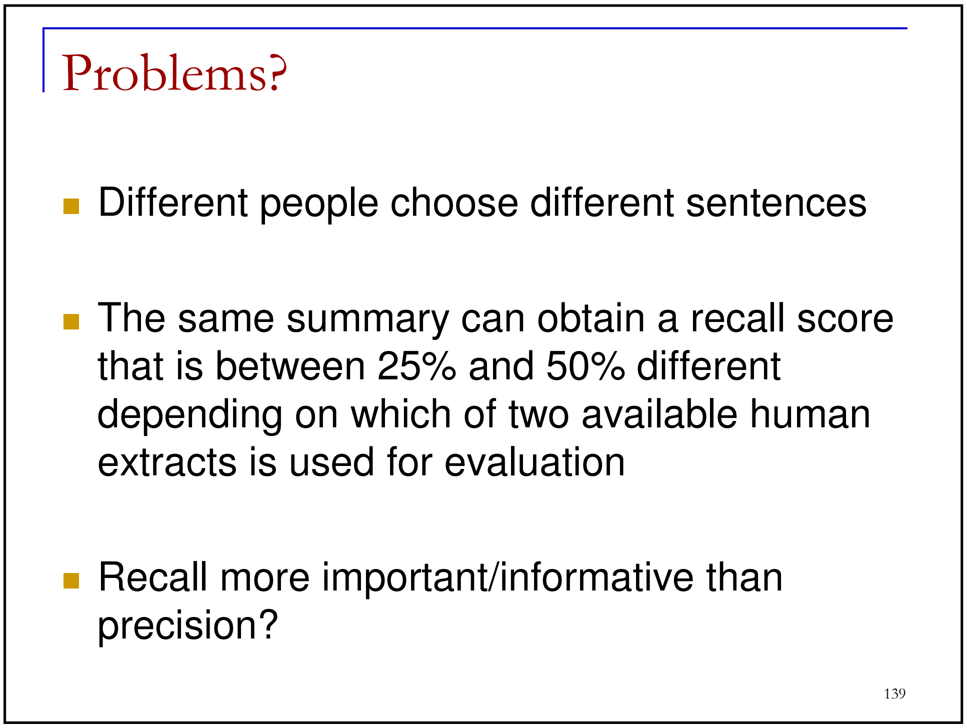
- No human gold-standard is used
  - Automatically compare input and summary

Precision and recall for extractive summaries

- Ask a person to select the most important sentences

Recall: system-human choice overlap/sentences chosen by human

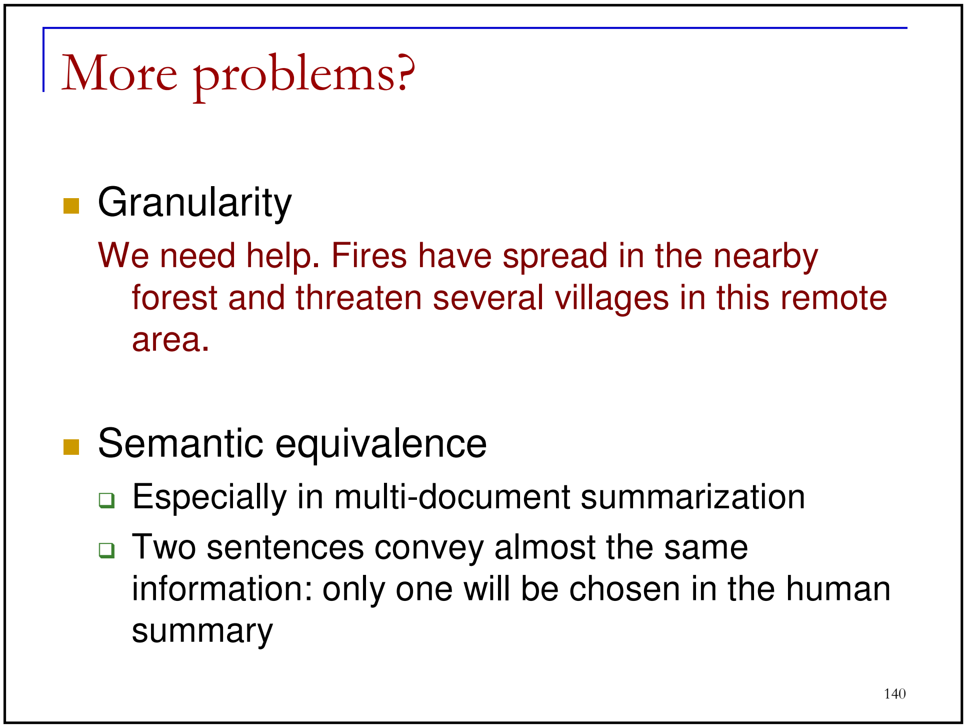
Precision: system-human choice overlap/sentences chosen by system



Problems?

- Different people choose different sentences
- The same summary can obtain a recall score that is between 25% and 50% different depending on which of two available human extracts is used for evaluation
- Recall more important/informative than precision?

139



More problems?

- Granularity

  We need help. Fires have spread in the nearby forest and threaten several villages in this remote area.

- Semantic equivalence
  - Especially in multi-document summarization
  - Two sentences convey almost the same information: only one will be chosen in the human summary

140

Evaluation methods for content

	Model summaries	Manual comparison/ ratings
Pyramid	✓	✓
Responsiveness	✗	✓
ROUGE	✓	✗
Fully automatic	✗	✗

Pyramid method [Nenkova and Passonneau, 2004; Nenkova et al., 2007]

- Based on Semantic Content Units (SCU)
- Emerge from the analysis of several texts
- Link different surface realizations with the same meaning

SCU example

S1 Pinochet arrested in London on Oct 16 at a Spanish judge's request for atrocities against Spaniards in Chile.

S2 Former Chilean dictator Augusto Pinochet has been arrested in London at the request of the Spanish government.

S3 Britain caused international controversy and Chilean turmoil by arresting former Chilean dictator Pinochet in London.

SCU: label, weight, contributors

Label London was where Pinochet was arrested

Weight=3

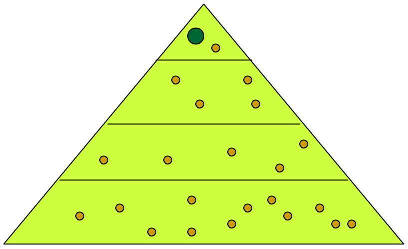
S1 Pinochet *arrested in London* on Oct 16 at a Spanish judge's request for atrocities against Spaniards in Chile.

S2 Former Chilean dictator Augusto Pinochet has been *arrested in London* at the request of the Spanish government.

S3 Britain caused international controversy and Chilean turmoil by arresting former Chilean dictator Pinochet *in London*.

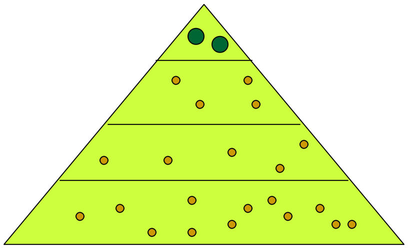
Ideally informative summary

- Does not include an SCU from a lower tier unless all SCUs from higher tiers are included as well



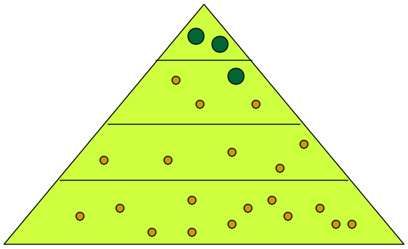
Ideally informative summary

- Does not include an SCU from a lower tier unless all SCUs from higher tiers are included as well



Ideally informative summary

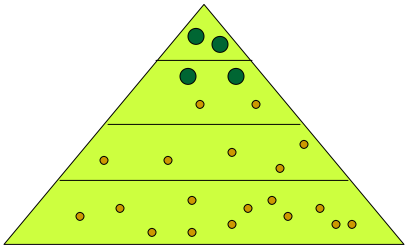
- Does not include an SCU from a lower tier unless all SCUs from higher tiers are included as well



147

Ideally informative summary

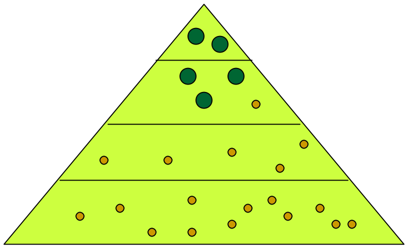
- Does not include an SCU from a lower tier unless all SCUs from higher tiers are included as well



148

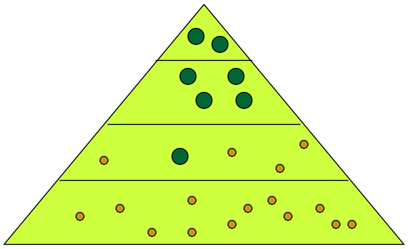
Ideally informative summary

- Does not include an SCU from a lower tier unless all SCUs from higher tiers are included as well



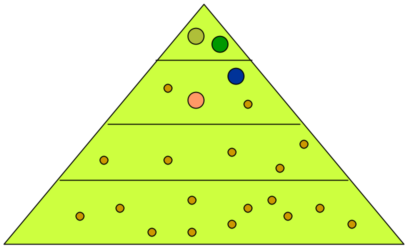
Ideally informative summary

- Does not include an SCU from a lower tier unless all SCUs from higher tiers are included as well



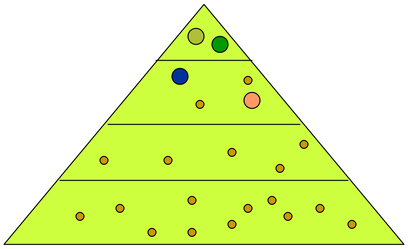
Different equally good summaries

- Pinochet arrested
- Arrest in London
- Pinochet is a former Chilean dictator
- Accused of atrocities against Spaniards



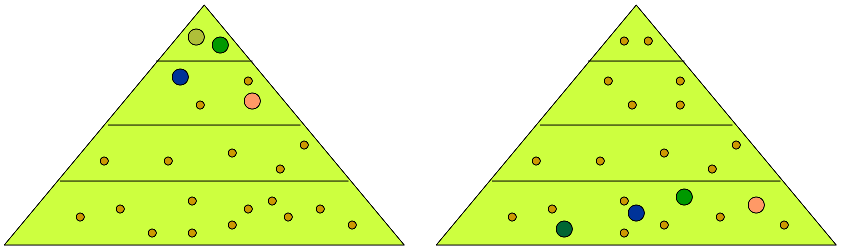
Different equally good summaries

- Pinochet arrested
- Arrest in London
- On Spanish warrant
- Chile protests



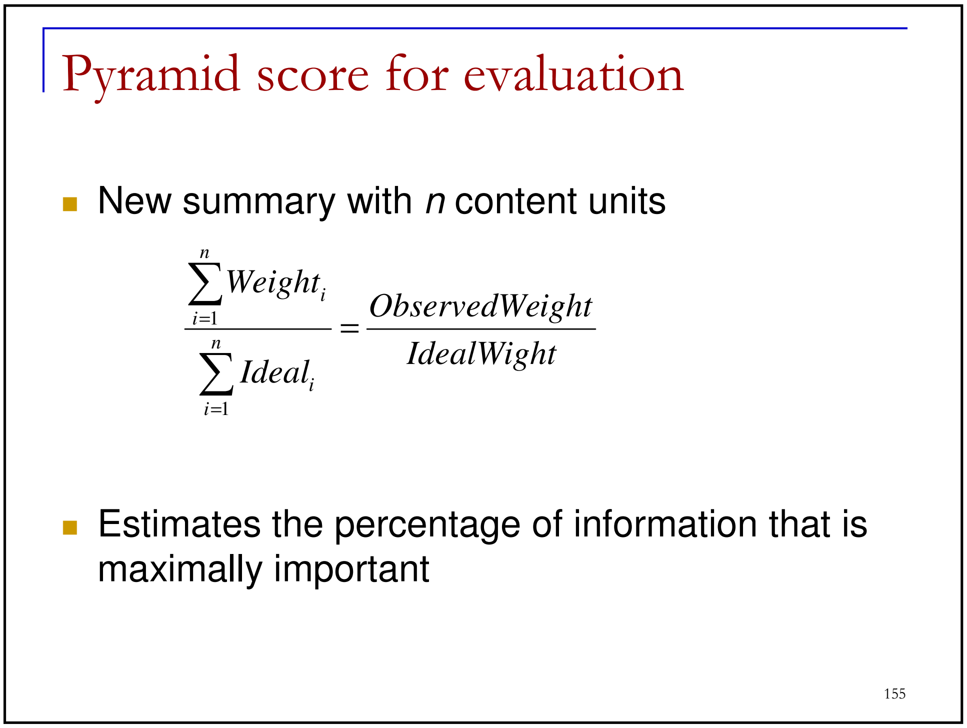
Diagnostic — why is a summary bad?

- Good
  - Less relevant summary



Importance of content

- Can observe distribution in human summaries
  - Assign relative importance
  - Empirical rather than subjective
- The more people agree, the more important

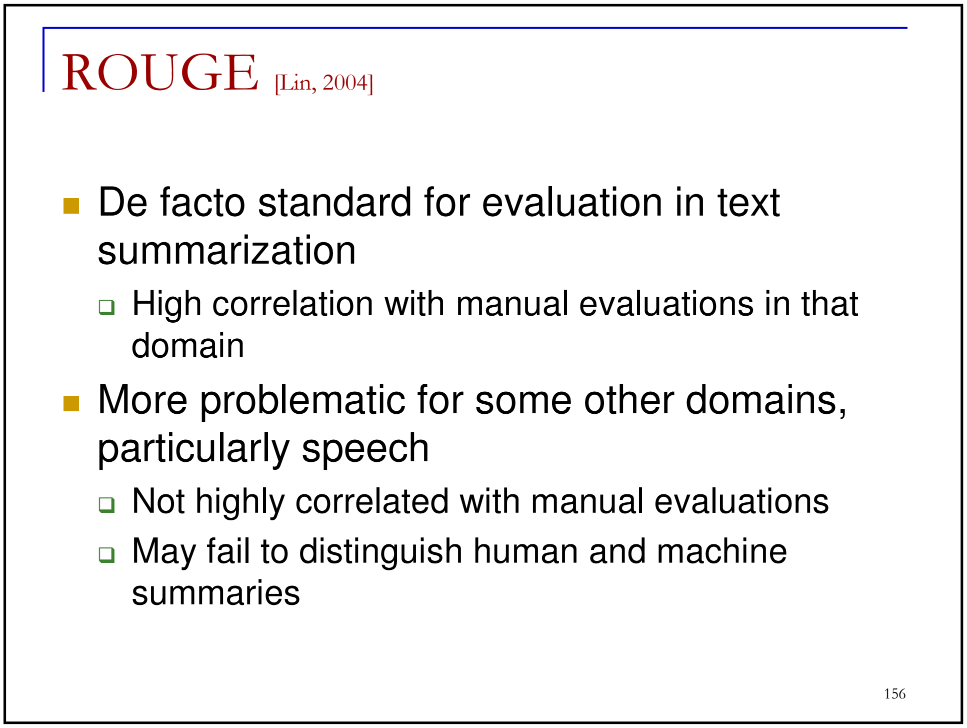


Pyramid score for evaluation

- New summary with *n* content units

$$\frac{\sum_{i=1}^{n} Weight_i}{\sum_{i=1}^{n} Ideal_i} = \frac{ObservedWeight}{IdealWight}$$

- Estimates the percentage of information that is maximally important



ROUGE [Lin, 2004]

- De facto standard for evaluation in text summarization
  - High correlation with manual evaluations in that domain
- More problematic for some other domains, particularly speech
  - Not highly correlated with manual evaluations
  - May fail to distinguish human and machine summaries

ROUGE details

- In fact a suite of evaluation metrics
  - Unigram
  - Bigram
  - Skip bigram
  - Longest common subsequence
- Many settings concerning
  - Stopwords
  - Stemming
  - Dealing with multiple models

How to evaluate without human involvement? [Louis and Nenkova, 2009]

- A good summary should be similar to the input
- Multiple ways to measure similarity
  - Cosine similarity
  - KL divergence
  - JS divergence
- Not all work!

JS divergence between input and summary

- Distance between two distributions as average KL divergence from their mean distribution

$$JS(Inp \,\|\, Summ) = \tfrac{1}{2}[\ KL(Inp \,\|\, A)\ +\ KL(Summ \,\|\, A)]$$

$$A = \frac{Inp + Summ}{2}\text{, } mean\ distribution\ of\ \ Input\ and\ Summary$$

Summary likelihood given the input

- Probability that summary is generated according to term distribution in the input

Higher likelihood ~ better summary

- Unigram Model $p_{Inp}(w_1)^{n_1}\, p_{Inp}(w_2)^{n_2} \ldots p_{Inp}(w_r)^{n_r}$

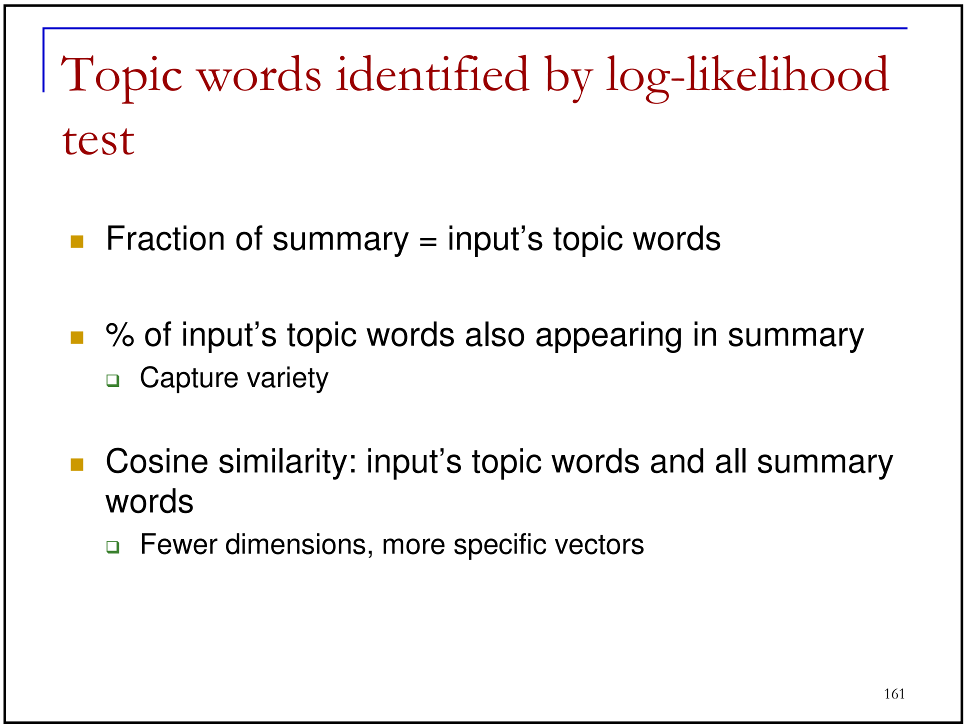
$$r - summary\ \ vocabulary$$

$$n_i = count\ in\ summary\ \ of\ \ word\ \ w_i$$

- Multinomial Model

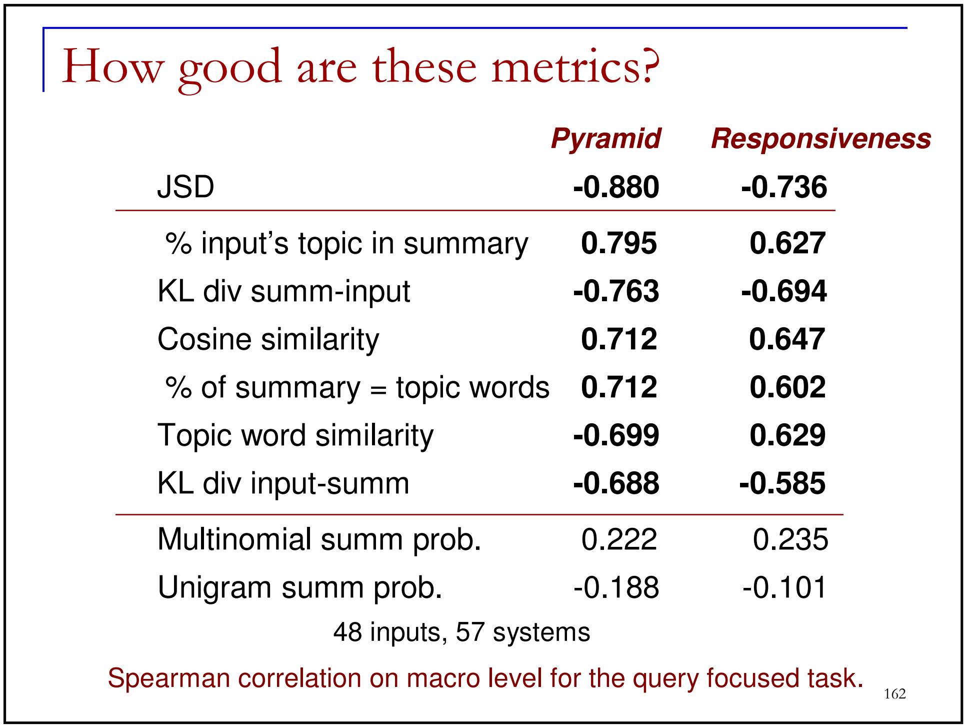
$$\tfrac{N!}{n_1! \ldots n_r!}\, p_{Inp}(w_1)^{n_1}\, p_{Inp}(w_2)^{n_2} \ldots p_{Inp}(w_r)^{n_r}$$

$$N = \sum_i n_i = summarysize$$



Topic words identified by log-likelihood test

- Fraction of summary = input’s topic words
- % of input’s topic words also appearing in summary
  - Capture variety
- Cosine similarity: input’s topic words and all summary words
  - Fewer dimensions, more specific vectors

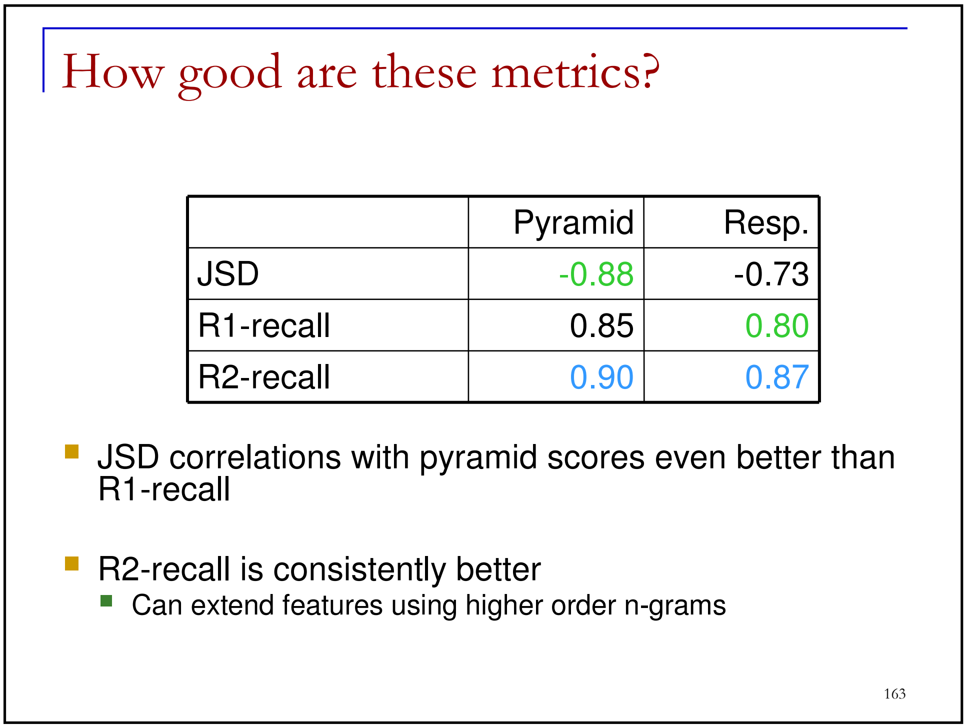


How good are these metrics?

	Pyramid	***Responsiveness***
JSD	**-0.880**	**-0.736**
% input’s topic in summary	**0.795**	**0.627**
KL div summ-input	**-0.763**	**-0.694**
Cosine similarity	**0.712**	**0.647**
% of summary = topic words	**0.712**	**0.602**
Topic word similarity	**-0.699**	**0.629**
KL div input-summ	**-0.688**	**-0.585**
Multinomial summ prob.	0.222	0.235
Unigram summ prob.	-0.188	-0.101

48 inputs, 57 systems

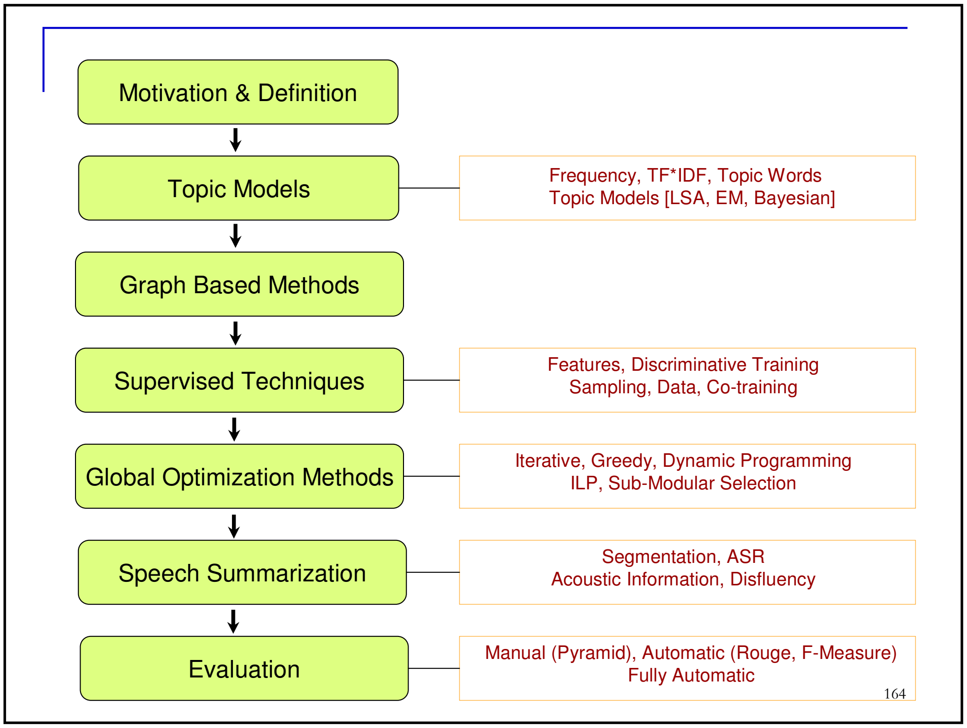
Spearman correlation on macro level for the query focused task.



How good are these metrics?

	Pyramid	Resp.
JSD	-0.88	-0.73
R1-recall	0.85	0.80
R2-recall	0.90	0.87

- JSD correlations with pyramid scores even better than R1-recall
- R2-recall is consistently better
  - Can extend features using higher order n-grams



Motivation & Definition

Topic Models — Frequency, TF*IDF, Topic Words
Topic Models [LSA, EM, Bayesian]

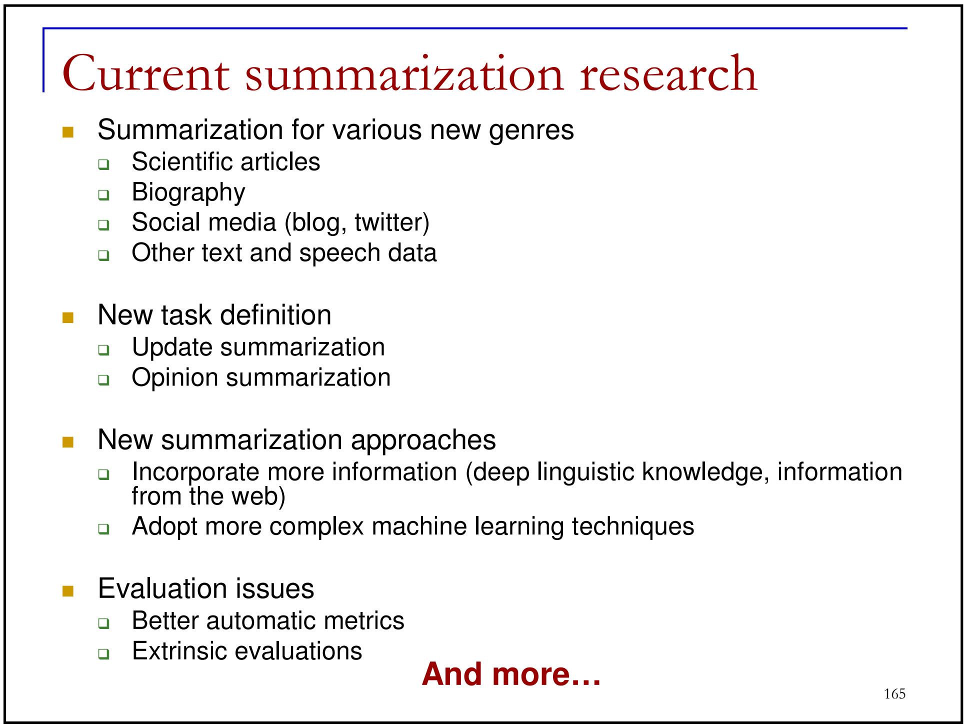
Graph Based Methods

Supervised Techniques — Features, Discriminative Training
Sampling, Data, Co-training

Global Optimization Methods — Iterative, Greedy, Dynamic Programming
ILP, Sub-Modular Selection

Speech Summarization — Segmentation, ASR
Acoustic Information, Disfluency

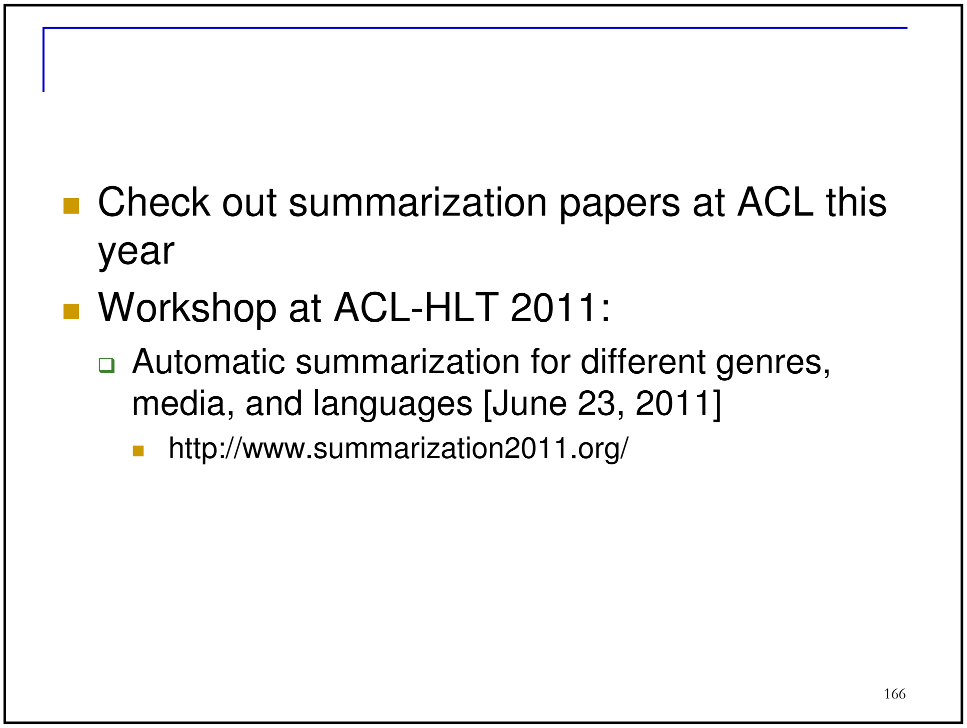
Evaluation — Manual (Pyramid), Automatic (Rouge, F-Measure)
Fully Automatic



Current summarization research

- Summarization for various new genres
  - Scientific articles
  - Biography
  - Social media (blog, twitter)
  - Other text and speech data
- New task definition
  - Update summarization
  - Opinion summarization
- New summarization approaches
  - Incorporate more information (deep linguistic knowledge, information from the web)
  - Adopt more complex machine learning techniques
- Evaluation issues
  - Better automatic metrics
  - Extrinsic evaluations

And more…



- Check out summarization papers at ACL this year
- Workshop at ACL-HLT 2011:
  - Automatic summarization for different genres, media, and languages [June 23, 2011]
    - http://www.summarization2011.org/

References

- Ahmet Aker, Trevor Cohn, Robert Gaizauska. 2010. Multi-document summarization using A* search and discriminative training. Proc. of EMNLP.
- R. Barzilay and M. Elhadad. 2009. Text summarizations with lexical chains. In: I. Mani and M. Maybury (eds.): Advances in Automatic Text Summarization.
- Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based reranking for Reordering Documents and Producing Summaries. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.
- H. Christensen, Y. Gotoh, B. Killuru, and S. Renals. 2003. Are Extractive Text Summarization Techniques Portable to Broadcast News? Proc. of ASRU.
- John Conroy and Dianne O'Leary. 2001. Text Summarization via Hidden Markov Models. Proc. of SIGIR.
- J. M. Conroy, J. D. Schlesinger, and D. P. OLeary. 2006. Topic-Focused Multi-Document Summarization Using an Approximate Oracle Score. Proc. COLING/ACL 2006. pp. 152-159.
- Thomas Cormen, Charles E. Leiserson, and Ronald L. Rivest.1990. Introduction to algorithms. MIT Press.
- G. Erkan and D. R. Radev.2004. LexRank: Graph-based Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research (JAIR).
- Pascale Fung, Grace Ngai, and Percy Cheung. 2003. Combining optimal clustering and hidden Markov models for extractive summarization. Proceedings of ACL Workshop on Multilingual Summarization.Sadoki Furui, T. Kikuchi, Y. Shinnaka, and C. Hori. 2004. Speech-to-text and Speech-to-speech Summarization of Spontaneous Speech. IEEE Transactions on Audio, Speech, and Language Processing. 12(4), pages 401-408.
- Michel Galley. 2006. A Skip-Chain Conditional Random Field for Ranking Meeting Utterances by Importance. Proc. of EMNLP.
- Dan Gillick, Benoit Favre. 2009. A scalable global model for summarization. Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing.
- Dan Gillick, Korbinian Riedhammer, Benoit Favre, Dilek Hakkani-Tur. 2009. A global optimization framework for meeting summarization. Proceedings of ICASSP.
- Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization.

References

- Y. Gong and X. Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. Proc. ACM SIGIR.
- I. Gurevych and T. Nahnsen. 2005. Adapting Lexical Chaining to Summarize Conversational Dialogues. Proc. RANLP.
- B. Hachey, G. Murray, and D. Reitter.2006. Dimensionality reduction aids term co-occurrence based multi-document summarization. In: SumQA 06: Proceedings of the Workshop on Task-Focused Summarization and Question Answering.
- Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. Proc. of NAACL-HLT.
- L. He, E. Sanocki, A. Gupta, and J. Grudin. 2000. Comparing presentation summaries: Slides vs. reading vs. listening. Proc. of SIGCHI on Human factors in computing systems.
- C. Hori and Sadaoki Furui. 2001. Advances in Automatic Speech Summarization. Proc. of Eurospeech.
- T. Kikuchi, S. Furui, and C. Hori. 2003. Automatic Speech Summarization based on Sentence Extractive and Compaction. Proc. of ICSLP.
- Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A Trainable Document Summarizer. Proc. of SIGIR.
- J. Leskovec, N. Milic-frayling, and M. Grobelnik. 2005. Impact of Linguistic Analysis on the Semantic Graph Coverage and Learning of Document Extracts. Proc. AAAI.
- Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries, Workshop on Text Summarization Branches Out.
- C.Y. Lin and E. Hovy. 2000. The automated acquisition of topic signatures for text summarization. Proc. COLING.
- Hui Lin and Jeff Bilmes. 2010. Multi-document summarization via budgeted maximization of submodular functions. Proc. of NAACL.
- Hui Lin and Jeff Bilmes and Shasha Xie. 2009. Graph-based Submodular Selection for Extractive Summarization. Proceedings of ASRU.
- Shih-Hsiang Lin and Berlin Chen. 2009. Improved Speech Summarization with Multiple-hypothesis Representations and Kullback-Leibler Divergence Measures. Proc. of Interspeech.
- Shih-Hsiang Lin, Berlin Chen, and H. Min Wang. 2009. A Comparative Study of Probabilistic Ranking Models for Chinese Spoken Document Summarization. ACM Transactions on Asian Language Information Processing.

References

- Shih Hsiang Lin, Yu Mei Chang, Jia Wen Liu, Berlin Chen. 2010 Leveraging Evaluation Metric-related Training Criteria for Speech Summarization. Proc. of ICASSP.
- Fei Liu and Yang Liu. 2009. From Extractive to Abstractive Meeting Summaries: Can it be done by sentence compression? Proc. of ACL.
- Fei Liu and Yang Liu. 2010. Using Spoken Utterance Compression for Meeting Summarization: A pilot study. Proc. of IEEE SLT.
- Yang Liu and Shasha Xie. 2008. Impact of Automatic Sentence Segmentation on Meeting Summarization. Proc. of ICASSP.
- Yang Liu, Feifan Liu, Bin Li, and Shasha Xie. 2007. Do Disfluencies Affect Meeting Summarization: A pilot study on the impact of disfluencies. Poster at MLMI.
- Yang Liu, Shasha Xie, and Fei Liu. 2010. Using n-best Recognition Output for Extractive Summarization and Keyword Extraction in Meeting Speech. Proc. of ICASSP.
- Annie Louis and Ani Nenkova. 2009. Automatically evaluating content selection in summarization without human models. Proceedings of EMNLP
- H.P. Luhn. 1958. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 2(2).
- Inderjeet Mani, Gary Klein, David House, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 2002. SUMMAC: a text summarization evaluation. Natual Language Engineering. 8,1 (March 2002), 43-68.
- Manuel J. Mana-Lopez, Manuel De Buenaga, and Jose M. Gomez-Hidalgo. 2004. Multidocument summarization: An added value to clustering in interactive retrieval. ACM Trans. Inf. Systems.
- Sameer Maskey. 2008. Automatic Broadcast News Summarization. Ph.D thesis. Columbia University.
- Sameer Maskey and Julia Hirschberg. 2005. Comparing lexical, acoustic/prosodic, discourse and structural features for speech summarization. Proceedings of Interspeech.
- Sameer Maskey and Julia Hirschberg. 2006. Summarizing Speech Without Text Using Hidden Markov Models. Proc. of HLT-NAACL.
- Ryan McDonald. 2007. A Study of Global Inference Algorithms in Multi-document Summarization. Lecture Notes in Computer Science. Advances in Information Retrieval.
- Kathleen McKeown, Rebecca J. Passonneau, David K. Elson, Ani Nenkova, and Julia Hirschberg. 2005. Do summaries help?. Proc. of SIGIR.
- K. McKeown, J. L. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin.1999. Towards multidocument summarization by reformulation: progress and prospects. Proc. AAAI 1999.

References

- R. Mihalcea and P. Tarau .2004. Textrank: Bringing order into texts. Proc. of EMNLP 2004.
- G. Murray, S. Renals, J. Carletta, J. Moore. 2005. Evaluating Automatic Summaries of Meeting Recordings. Proc. of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation.
- G. Murray, T. Kleinbauer, P. Poller, T. Becker, S. Renals, and J. Kilgour. 2009. Extrinsic Summarization Evaluation: A Decision Audit Task. ACM Transactions on Speech and Language Processing.
- A. Nenkova and R. Passonneau. 2004. Evaluating Content Selection in Summarization: The Pyramid Method. Proc. HLT-NAACL.
- A. Nenkova, L. Vanderwende, and K. McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. Proc. ACM SIGIR.
- A. Nenkova, R. Passonneau, and K. McKeown. 2007. The Pyramid Method: Incorporating human content selection variation in summarization evaluation. ACM Trans. Speech Lang. Processing.
- Miles Osborne. 2002. Using maximum entropy for sentence extraction. Proc. of ACL Workshop on Automatic Summarization.
- Gerald Penn and Xiaodan Zhu. 2008. A critical Reassessement of Evaluation Baselines for Speech Summarization. Proc. of ACL-HLT.
- Dmitri G. Roussinov and Hsinchun Chen. 2001. Information navigation on the web by clustering and summarizing query results. Inf. Process. Manage. 37, 6 (October 2001), 789-816.
- B. Schiffman, A. Nenkova, and K. McKeown. 2002. Experiments in Multidocument Summarization. Proc. HLT.
- A. Siddharthan, A. Nenkova, and K. Mckeown.2004. Syntactic Simplification for Improving Content Selection in Multi-Document Summarization. Proc. COLING.
- H. Grogory Silber and Kathleen F. McCoy. 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. Computational. Linguist. 28, 4 (December 2002), 487-496.
- J. Steinberger, M. Poesio, M. A. Kabadjov, and K. Jeek. 2007. Two uses of anaphora resolution in summarization. Inf. Process. Manage. 43(6).
- S. Tucker and S. Whittaker. 2008. Temporal compression of speech: an evaluation. IEEE Transactions on Audio, Speech and Language Processing, pages 790-796.
- L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova. 2007. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. Information Processing and Management 43.

References

- Kam-Fai Wong, Mingli Wu, and Wenjie Li. 2008. Extractive Summarization using Supervised and Semi-supervised learning. Proc. of ACL.
- Shasha Xie and Yang Liu. 2010. Improving Supervised Learning for Meeting Summarization using Sampling and Regression. Computer Speech and Language. V24, pages 495-514.
- Shasha Xie and Yang Liu. 2010. Using Confusion Networks for Speech Summarization. Proc. of NAACL.
- Shasha Xie, Dilek Hakkani-Tur, Benoit Favre, and Yang Liu. 2009. Integrating Prosodic Features in Extractive Meeting Summarization. Proc. of ASRU.
- Shasha Xie, Hui Lin, and Yang Liu. 2010. Semi-supervised Extractive Speech Summarization via Co-training Algorithm. Proc. of Interspeech.
- S. Ye, T.-S. Chua, M.-Y. Kan, and L. Qiu. 2007. Document concept lattice for text understanding and summarization. Information Processing and Management 43(6).
- W. Yih, J. Goodman, L. Vanderwende, and H. Suzuki. 2007. Multi-Document Summarization by Maximizing Informative Content-Words. Proc. IJCAI 2007.
- Klaus Zechner. 2002. Automatic Summarization of Open-domain Multiparty Dialogues in Diverse Genres. Computational Linguistics. V28, pages 447-485.
- Klaus Zechner and Alex Waibel. 2000. Minimizing word error rate in textual summaries of spoken language. Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference.
- Justin Zhang and Pascale Fung. 2009. Extractive Speech Summarization by Active Learning. Proc. of ASRU.
- Xiaodan Zhu and Gerald Penn. 2006. Comparing the Roles of Textual, Acoustic and Spoken-language Features on Spontaneous Conversation Summarization. Proc. of HLT-NAACL.
- Xiaodan Zhu, Gerald Penn, and F. Rudzicz. 2009. Summarizing Multiple Spoken Documents: Finding Evidence from Untranscribed Audio. Proc. of ACL.