Welcome to CIS 530/430

Can machines understand language?

- Which aspects of language are easy for people but difficult for machines?

Syntactic and semantic ambiguities

I ate cherry pie with ice cream.

I ate cherry pie with a spoon.

The astronomer married the star.

Jack called John at 6am.

He woke him up.

He got quite mad at him.

NLP in NYT
some noteable achievements

- [Computer Wins On 'Jeopardy!': Trivial, It's Not]
- [Aiming to Learn as We Do, a Machine Teaches Itself]
- [Speech Recognition's Early Days]
- [Google's Computing Power Refines Translation Tool]
- [A News Cocktail Mixed by a Software Genie]

In CIS 540/430 we will

- Introduce some of the classical problems addressed in NLP
- Broad overview, with emphasis on practical solutions
- Learn to soundly address empirical problems
  - Is one system for a task better than another
  - Understand where and how a system fails
  - Propose possible solutions
- Talk/write clearly about your work, decisions and observations

Topics

- Part of speech tagging
- Syntactic Parsing
- Automatic Speech Recognition
- The Representation of Meaning
- Computational Semantics
- Lexical Semantics
- Computational Lexical Semantics
- Computational Discourse
- Information Extraction
- Question Answering and Summarization
- Dialog and Conversational Agents
- Machine Translation

Textbooks

[Required] Natural Language Processing with Python---Analyzing Text with the Natural Language Toolkit, by Steven Bird, Ewan Klein, and Edward Loper.

- Available online http://www.nltk.org/book

[Optional] Speech and Language Processing (second edition, 2008, Prentice-Hall), by Daniel Jurafsky and James Martin http://www.cis.upenn.edu/~cis530/

Will be up by the end of the week

Slides, assigned readings and extra info will be posted there

Staff

- Instructor: Ani Nenkova
  - Levine 505, nenkova@seas.upenn.edu
  - Office hours: Wed, 4:30—5:30

- Teaching assistant: Alexander Shoulson
  - shoulson@seas.upenn.edu
  - Office hours: TBD

Grading

- 4 homework assignments: 60%
  - Implementation (Python)
  - Approach selection
  - Results analysis
- Class project: 35%
  - This year: automatic summarization
  - Competition at the end of the academic year
  - Phd students can submit alternative project proposal
- Class participation: 5%

Teamwork

- For projects students can work in teams of upto 3 people

- Academic integrity

- Questions?

Why summarize?

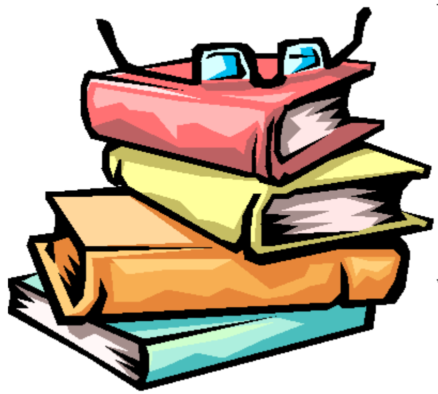
Text summarization

News articles

Emails

Social Media Streams

Scientific Articles



Books

Websites

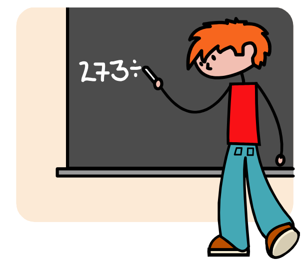
Speech summarization

Phone Conversation

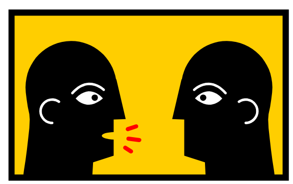
Lecture

Meeting

Talk Shows



Classroom



Chat

Broadcast News

Radio News

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]

Tasks in summarization

Content (sentence) selection

- Extractive summarization

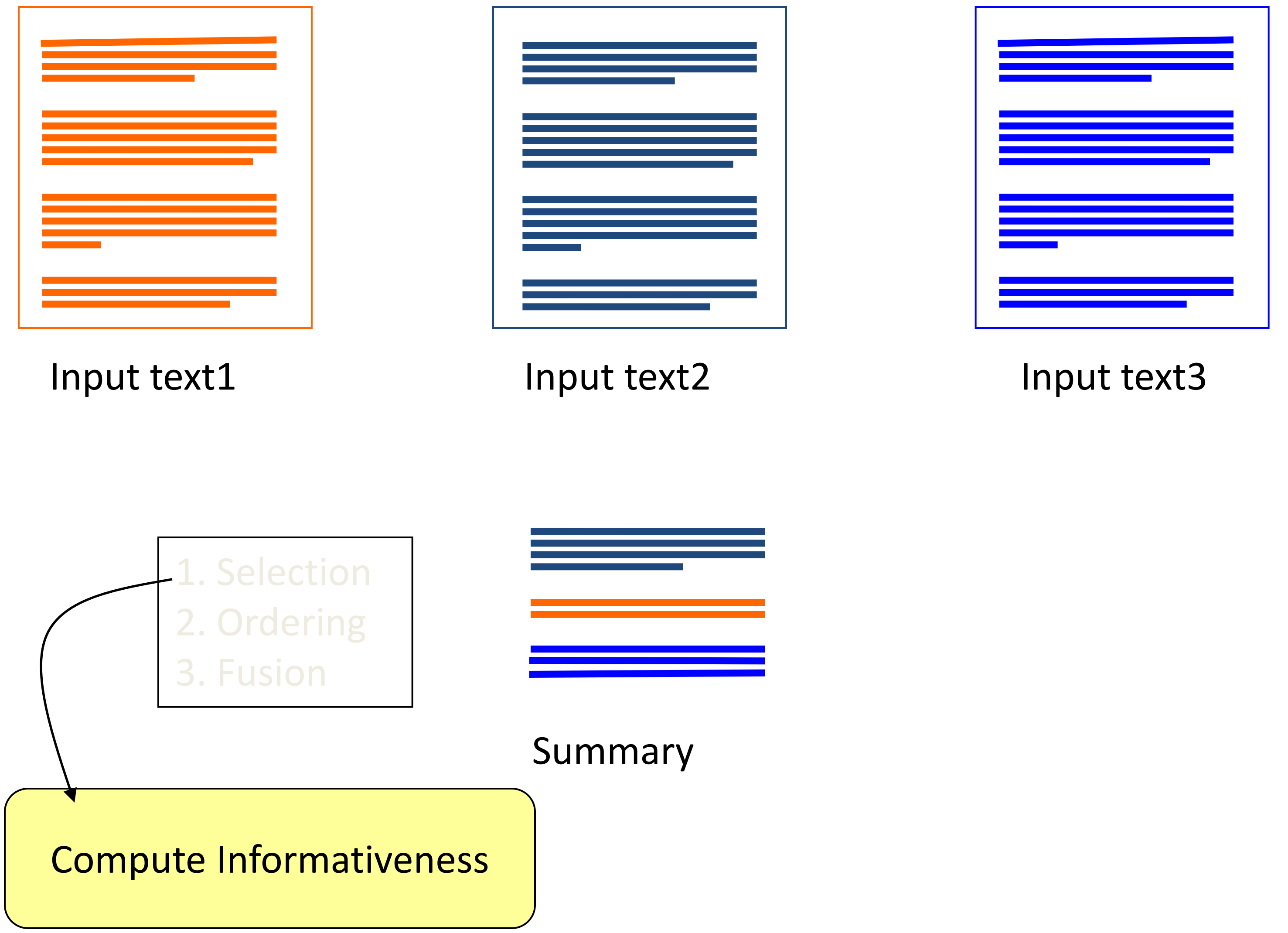
Information ordering

- 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



Steps in extractive summarization

- Derive a representation of the input
  - Topic representation: interpretable as aboutness
  - Indicator representation: indicators of importance
- Compute sentence scores (importance)
  - Gives a ranking of sentences
  - Based on this score, choose the summary sentences
- Select the summary
  - Greedy approaches that possibly take previous decisions in account
  - Optimize the set of selected sentences

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

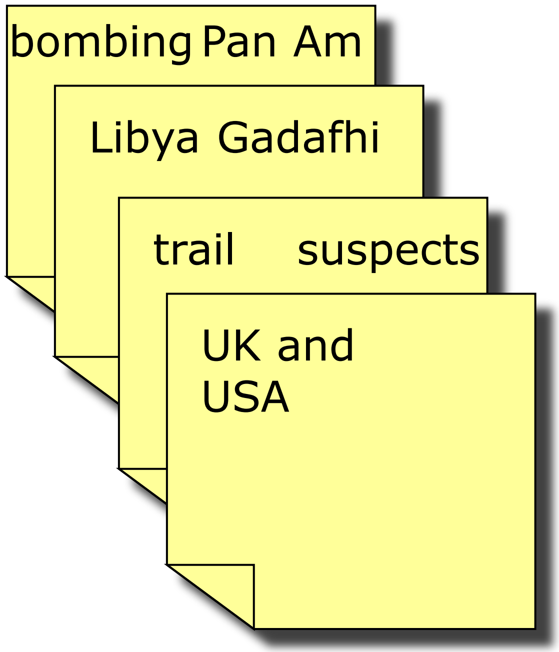
What is an article about?

- 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

INPUT



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

SUMMARY

Libya refuses to surrender two Pan Am bombing suspects

HOW?

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

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

- Update word probabilities for the selected sentence to reduce redundancy

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

- 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

HSSSSSSSSSHSSSHSSHHSHHHHH

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.

Document centroid as representation of document topic

- Consider all words in the document, $t_1, t_2, \ldots, t_n$
- Each sentence in the document is represented as a vector $\vec{s}_i = (w_{i1}, w_{i2}, \ldots, w_{in})$
  - ➢ If a word does not appear in the sentence, $w_{ij} = 0$
  - ➢ Else $w_{ij} = tf.idf_j$
- Centroid = average of all sentence vectors
  - – A table of words and their weights
  - – Find the sentence most similar to the centroid to produce the summary

Centroid

	t1	**t2**	**…**	**tn**
s1	w11	w12	…	w1n
s2	w21	w22	…	w2n
…	…	…	…	…
sk	wk1	wk2	…	wkn
centroid	$\sum_{j=1}^{k} w_{1j} / k$	$\sum_{j=1}^{k} w_{2j} / k$		$\sum_{j=1}^{k} w_{nj} / k$

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 input 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 (or weight)
  - Use standard change of coordinate system and dimensionality reduction techniques
  - In the new space each column of U corresponds to the most important topics in the corpus; P gives the weights of topics; V represents the sentences in terms of coverage of the topics
  - Select the best sentence to cover each topic

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

Location, time

(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.

damage

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.

...

magnitude

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.

...

relief efforts

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]

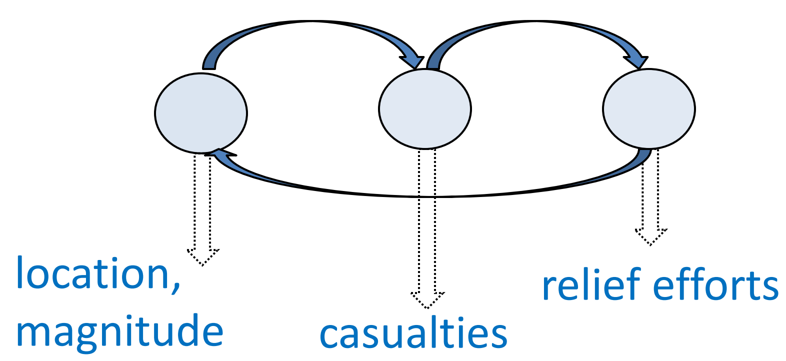
- Hidden Markov Model (HMM)-based
  - States - clusters of related sentences “topics”
  - Transition prob. - sentence precedence in corpus
  - Emission prob. - bigram language model

Generating sentence in current *topic*

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

Transition from previous *topic*

Earthquake reports



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.

KL divergence

- Distance between two probability distributions: P, Q

$$KL(P \parallel Q) = \sum_{w} p_P(w) \log_2 \frac{p_P(w)}{p_Q(w)}$$

- P, Q: Input and summary word distributions

$$KL(P \| Q) = \sum_w \log \frac{P(w)}{Q(w)}$$

KL divergence has become the most successful function for sentence scoring

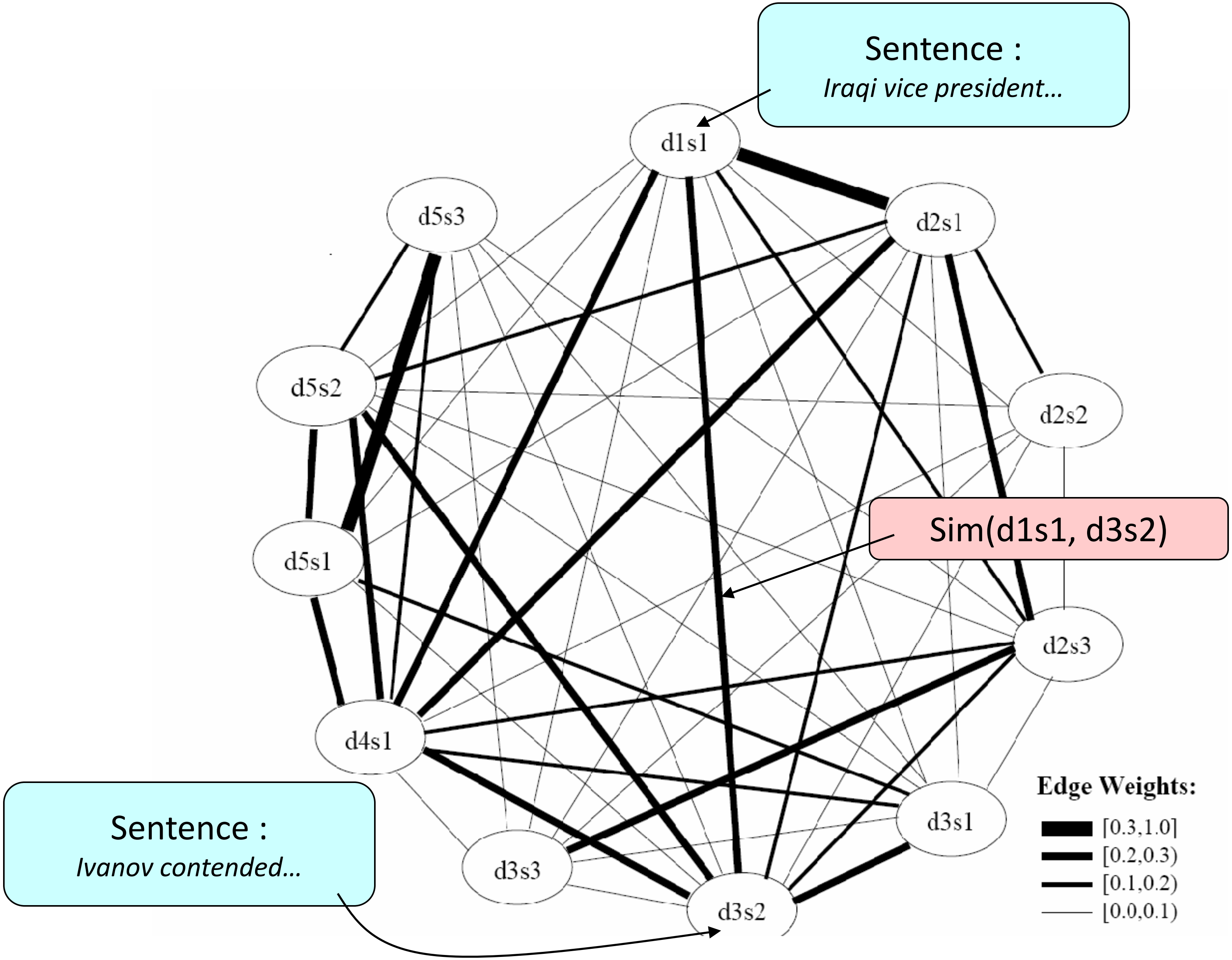
It has been used in Bayesian methods, in summarization of scientific articles and in evaluation

- 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…
Sentence :
Ivanov contended…
Sim(d1s1, d3s2)
d1s1
d2s1
d2s2
d2s3
d3s1
d3s2
d3s3
d4s1
d5s1
d5s2
d5s3
Edge Weights:
[0.3,1.0]
[0.2,0.3)
[0.1,0.2)
[0.0,0.1)

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
  - formulating summarization as a classification problem gives much flexibility, there is no need to choose a single sentence score

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

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

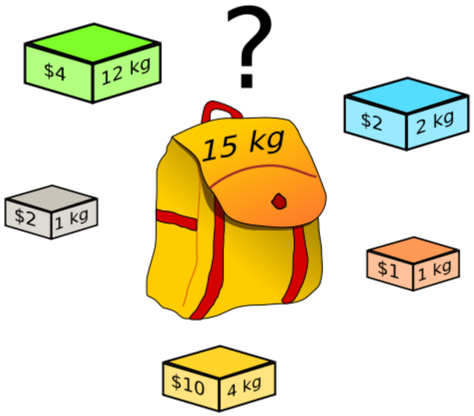
Summarization as an optimization problem

- 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

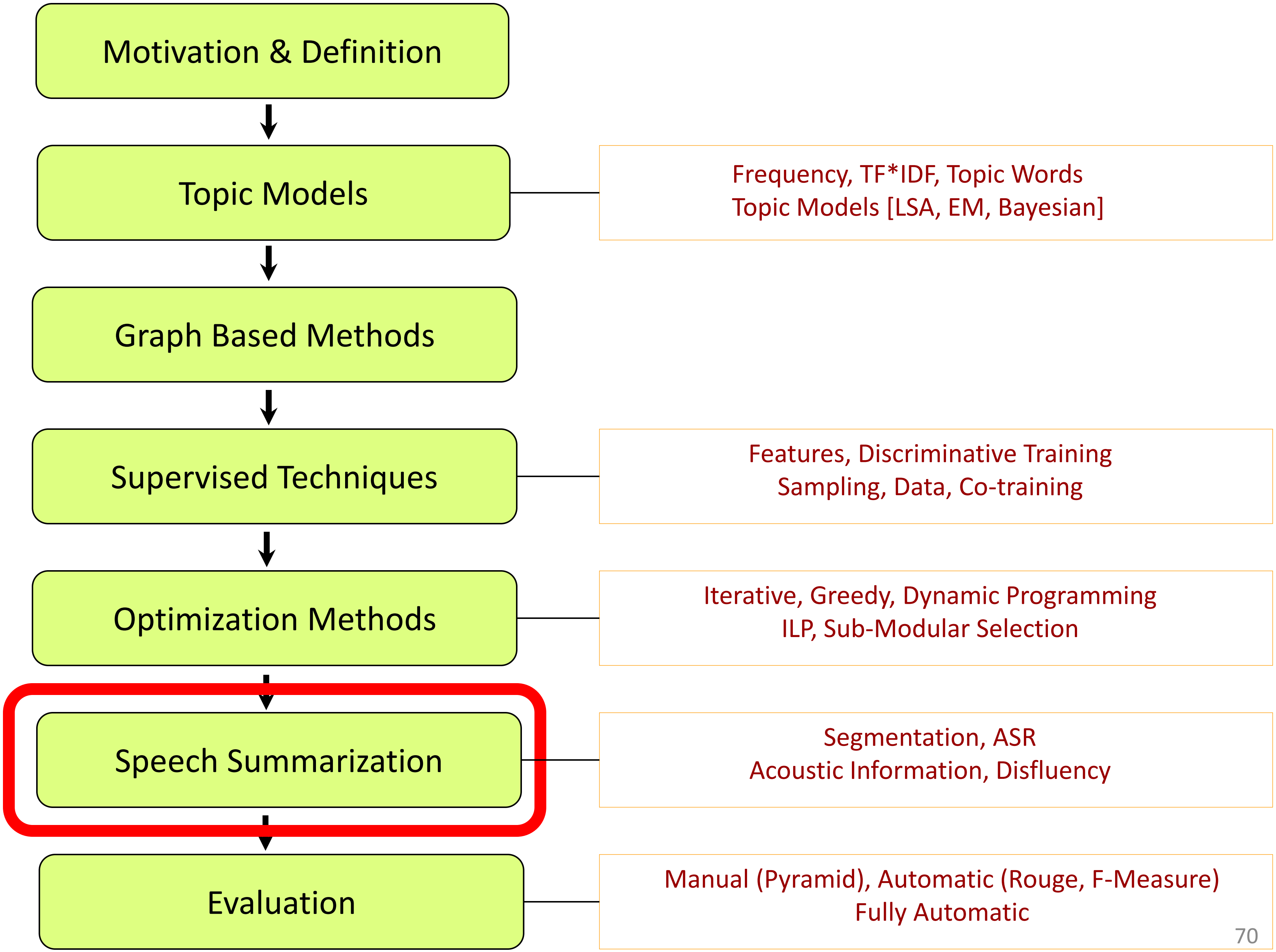


- Many other similar optimization problems
- General Idea: Maximize a function given a set of constraints

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
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

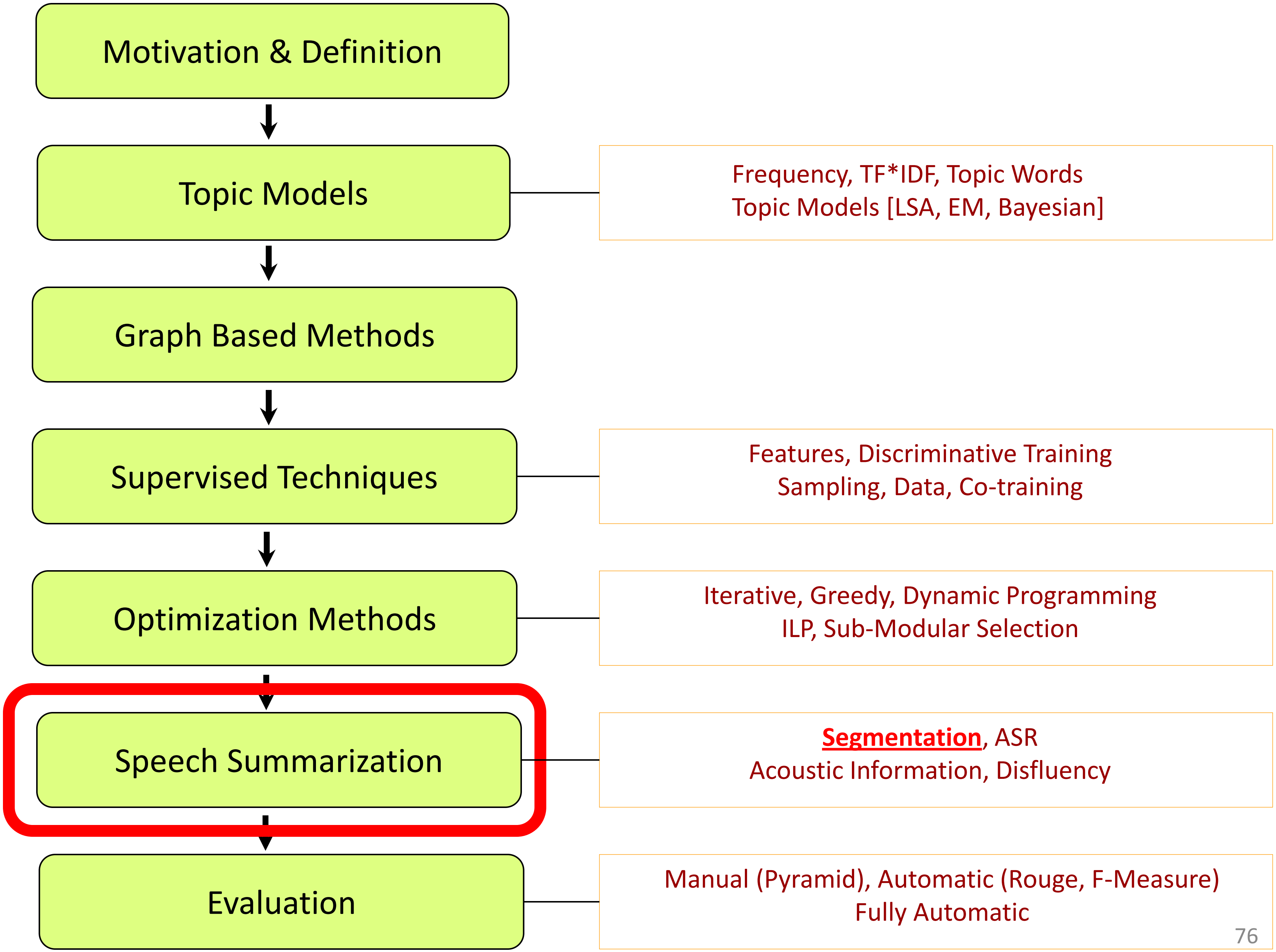
try to use electrical appliances before p.m. and after p.m. and turn off computers, copiers and lights when they're not being used

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]

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)

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