

# Content selection in extractive summarization

Lecture 13

Cis 530/430

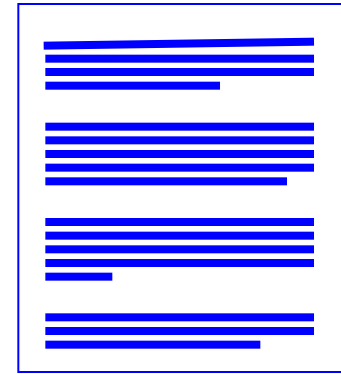
# Extractive (multi-document) summarization



Input text1

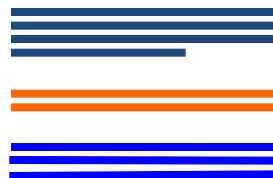


Input text2



Input text3

1. Selection
2. Ordering
3. Fusion



Summary

Compute Informativeness

# 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.

# Log-likelihood test (topic signature terms)

- $t$ : term
- $T$ : cluster of articles on a given topic
- $NT$ : set of articles not on that topic

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)

# How to decide which hypothesis is more likely?

- Consider a text to be a sequence of Bernoulli trials
  - A word is either our term of interest  $t$  or not
  - The likelihood of observing term  $\underline{t}$  which occurs with probability  $\underline{p}$  in a text consisting of  $\underline{N}$  words is given by

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

# Log-likelihood ratio

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

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

Statistical significance can be computed at any desired level. At that 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

# Sentence clustering for theme identification

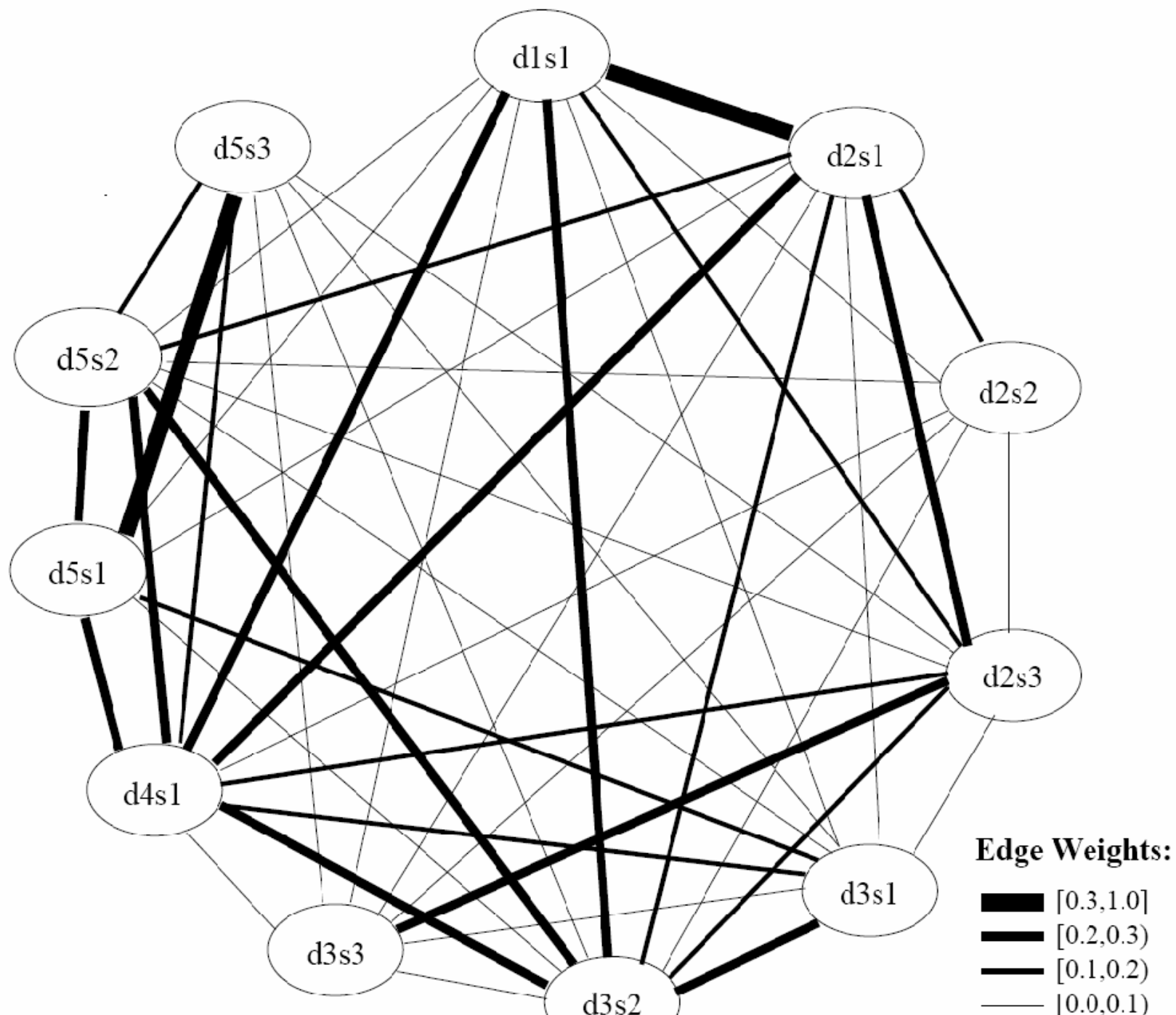
1. PAL was devastated by a pilots' strike in June and by the region's currency crisis.
2. In June, PAL was embroiled in a crippling three-week pilots' strike.
3. Tan wants to retain the 200 pilots because they stood by him when the majority of PAL's pilots staged a devastating strike in June.

→ Choose one sentence to represent each cluster

# Using graph representations

- 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



# 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
  - The probability of being in each node can be computed (stationary distribution)

# PageRank

- Probability of being at a given node in a graph if randomly following paths according to the probabilities defined in the graph
  - Need to normalize the similarities to turn them into probabilities
  - A node-to-node similarity matrix is transformed to a matrix of conditional probability of transitioning from one node to another

# Centrality

- The probability of each node depends on the probability of being at an adjacent node in the graph

$$p(u) = \sum_{v \in \text{adj}[u]} \frac{p(v)}{\text{deg}(v)}$$

- Or each node can be chosen with uniform probability  $\frac{1}{N}$
- Combine both as a weighted combination ( $d=0.85$ )

$$p(u) = (1-d) \frac{1}{N} + d \sum_{v \in \text{adj}[u]} \frac{p(v)}{\text{deg}(v)}$$

# Iterative approximation

- Stationary distribution: probability of being at each node after a long time of random walk
  - Guaranteed to exist for walks defined by the last equation
- Can be computed iteratively; fast convergence

$$1) \quad p(n_i, 0) = \frac{1}{N}$$

$$2) \quad p(n_i, t+1) = (1-d) \frac{1}{N} + d \sum_{n_j \in \text{ag}(n_i)} \frac{p(n_j, t)}{\text{deg}(n_j)}$$

# Various possibilities

- Continuous similarity vs. binary based on a threshold
  - Binary gives better results for content selection
- Is the degree of a node in the binary graph of any use?
  - In the binary threshold graph, the results from using LexRank and simply the degree of the node to weight the node are the same

# Summarization of blogs and academic papers

- Exploiting information about what others say about the text we need to summarize
  - An important segment of a blog is one that generates many comments
  - An important aspect of an academic paper is one that other papers citing the target one talk about

# Idea

- Estimate a language model from
  - the comments to the blog
  - Sentences in other papers that cite a given paper
- Then from the original blog or paper, select the sentences that are most likely generated by that language model

$$w(s) = P(w_1 w_2 \dots w_n)^{\frac{-1}{N}}$$

# Alternatively

- Consider that each sentence defines a probability distribution over words
  - Estimate this distribution
  - Will be sparse, so smoothing will be necessary
- Find the sentences in the original document that define distributions that differ least from the LM estimated from comments
  - How to compute similarities between distributions

# 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: sentence and comment distributions word distributions

# KL in the basic summarization model

Start with an empty summary SUM

**Step 1** Estimate word weights (probabilities)

**Step 2** Estimate sentence weights

$$\textit{Weight}(S) = \textit{KL}(\textit{Input} \mid \textit{SUM} \cup S)$$

**Step 3** Choose best sentence

**Step 4** Update word weights

**Step 5** Go to 2 if desired length not reached

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

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

# Project ideas

- Many existing summarizers have been introduced as a complete package of
  - Sentence representation
  - Sentence scoring
  - Sentence selection
- Figuring out which component contributes most to the system performance would be useful
- As well as combining promising aspects from different systems
  - Can one use topic words but do scoring by KL?

# Single document summarization in a cluster of documents

- This setting is similar to the blog-comment and article-citation
  - Can the language model approach be used to figure out which content is related to the overall topic of the cluster and which is specific to that document
  - Can you come up with an approach to figure out what content is specific to a document AND important (hard)

# 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
  - Nice because one does not need to commit to a specific topic representation!
- Select the sentences with highest scores

# 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

# Other questions to address in a supervised manner

- Instead of asking which sentences should be in the summary, can we predict which words should be in the summary
- And how many times they should appear?

- Next lecture we will talk about evaluation
- This will complete our preparation for project proposals
- Check out the project resources on the class webpage