

Content selection in extractive summarization

Lecture 13

Cis 530/430

Project proposals

due Nov 10!

Each proposal should contain the four sections listed below. It shouldn't be more than 2 pages of plain text. Shorter is fine if you have addressed all the points.

- General motivation
- Specific task to be addressed
- Approach and resources
- Evaluation

General motivation

- Long-term goal or application
 - A system that performs analysis of news to aid investment decisions
 - A system that is capable of generating new content based on textual input (synthesis or inference)
 - Choose the best compromise between resource need, speed and accuracy to select a summarizer for my company
 - Analyze the differences between consensus and coverage summarization
 - Simplify summaries for less proficient readers

Gives context to your project, but will not be attained at the completion of your project

Specific task to be addressed

- Your first step towards achieving your big vision. Much smaller, exciting probably only as seen as a first step towards your big task
 - It has to be concrete enough that you can say which part of it you will do tomorrow and what you plan to do next week
 - Implement algorithm X
 - Find all words that express positive or negative sentiment in a text (then next week work on an algorithm for summarizing such sentences)

Approach and resources

- Will your method use machine learning or will it be rule-based?
- Will you use a parser? A POS tagger? Any other tool?
- Do you need any specific type of data, annotated or not? How will you get these data?
- What will you implement and what functionality do you plan to get from existing tools

You should incorporate 2-3 existing tools in your system!

Evaluation

- What is the exact output you expect from your system or study?
- How will you measure the quality of the output?
- What is your baseline (simple approach, or existing practice)
- How will you perform the evaluation?

Project timeline and related work

- For your own sanity, it will be helpful to actually get a realistic timeline of what part of the project will be completed by what time! Take into account other exams and coursework, holidays etc
- Find related papers. They will give you ideas about specific experiments you'd like to perform, and will be the basis of your literature review.

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?

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

DUC/TAC

<http://duc.nist.gov/>

<http://www.nist.gov/tac/>

- Most systems have been developed and evaluated on data from the Text Analysis Conferences (TAC)
 - Single document summarization
 - Generic multi-document
 - Query-focused multi-document
 - Update
 - Headline generation
- Many systems evaluated manually and automatically on the same data

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?

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

Evaluation methods for content

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

Pyramid method

- 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

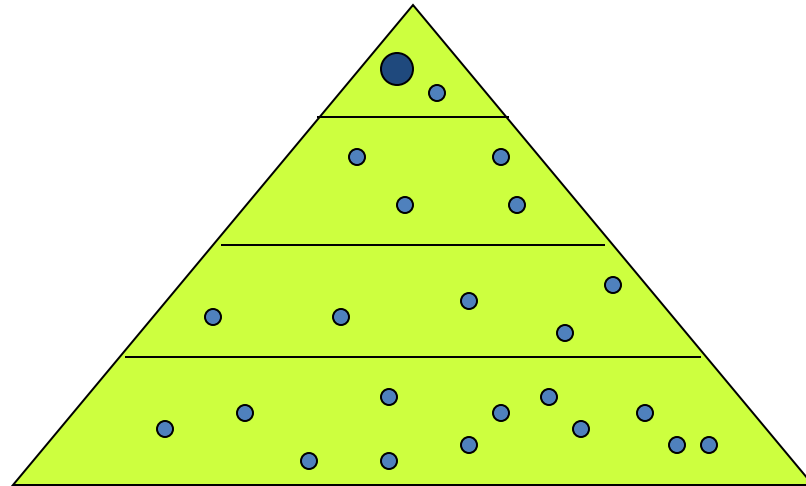
Label London was where Pinochet was arrested

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*.

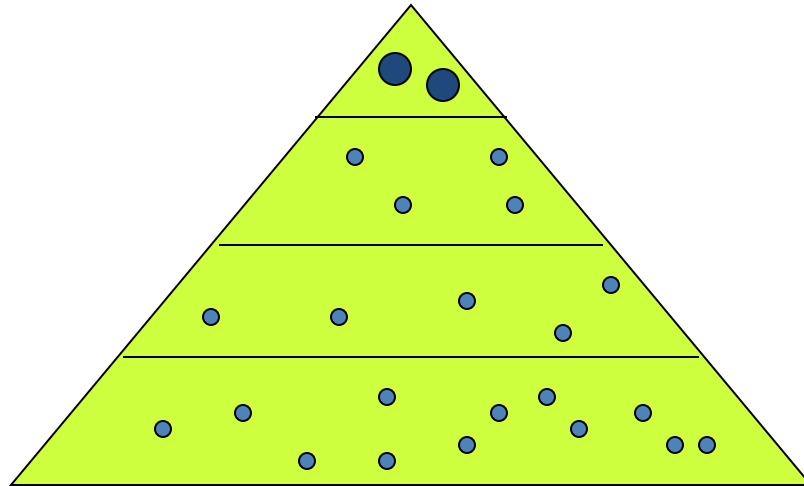
Ideally informative summary

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



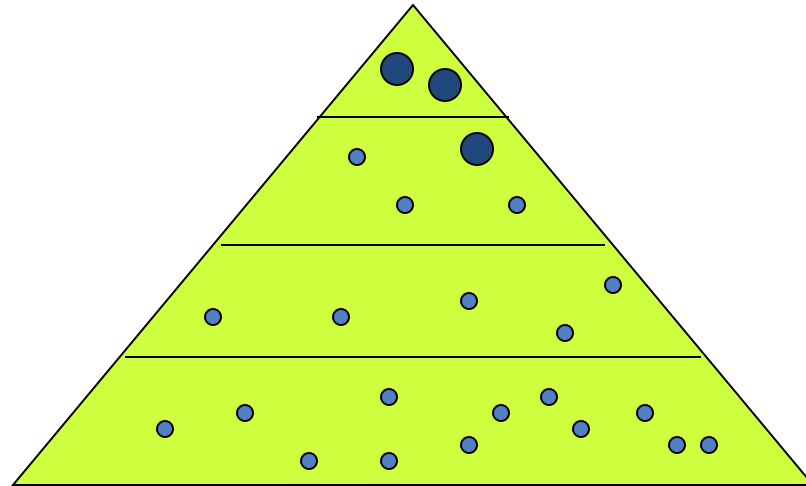
Ideally informative summary

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



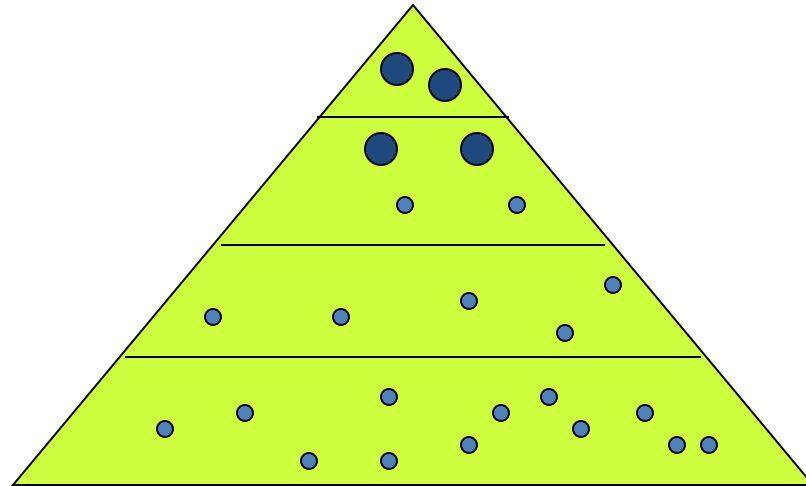
Ideally informative summary

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



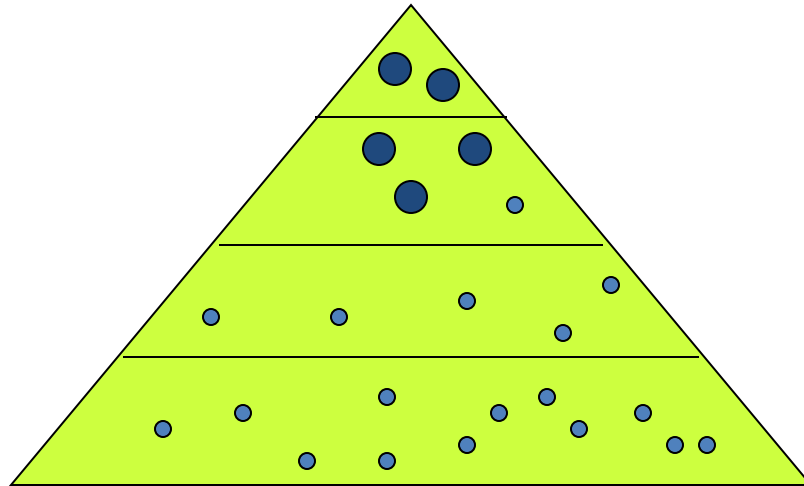
Ideally informative summary

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



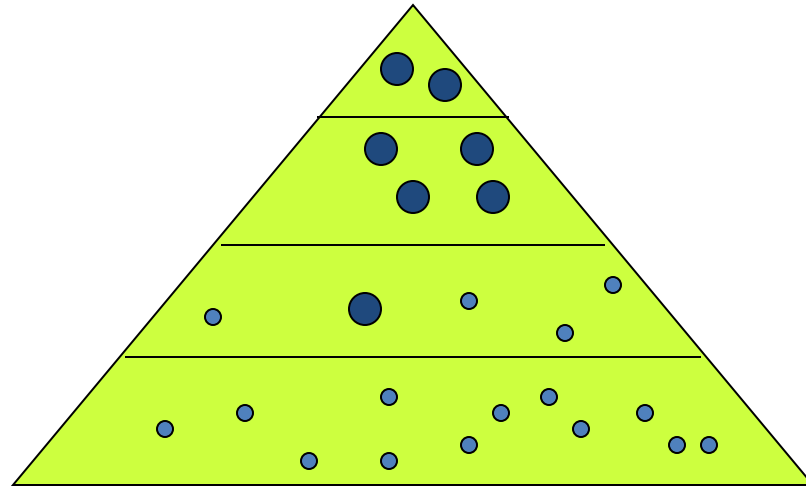
Ideally informative summary

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



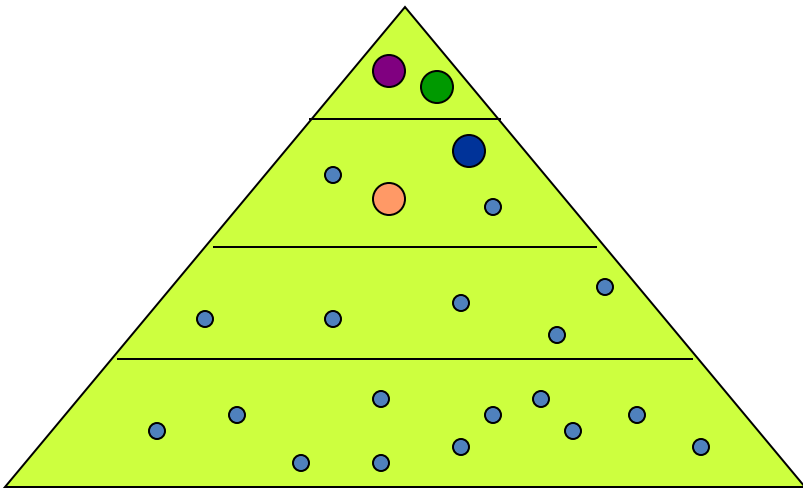
Ideally informative summary

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



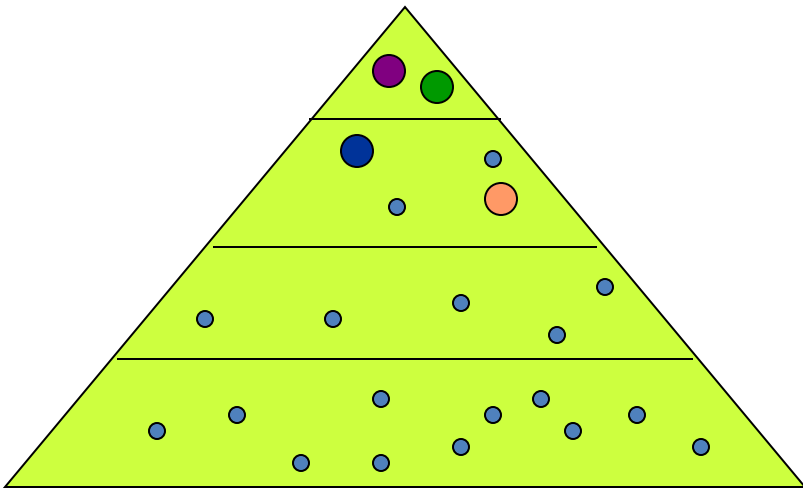
Different equally good summaries

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



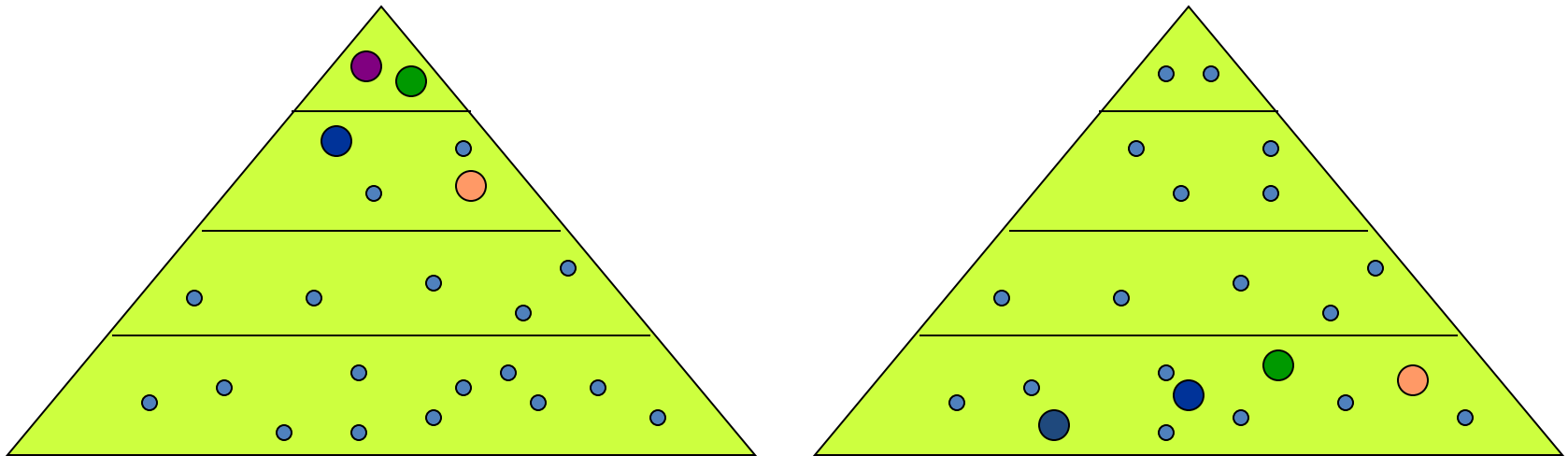
Different equally good summaries

- 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

$$JS(Inp \parallel Summ) = \frac{1}{2} [KL(Inp \parallel A) + KL(Summ \parallel A)]$$

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

Summary likelihood given the input

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

Higher likelihood ~ better summary

- Unigram Model $P_{Inp}(w_1)^{n_1} P_{Inp}(w_2)^{n_2} \dots P_{Inp}(w_r)^{n_r}$

r – summary vocabulary

n_i = count in summary of word w_i

- Multinomial Model

$$\frac{N!}{n_1! \dots n_r!} P_{Inp}(w_1)^{n_1} P_{Inp}(w_2)^{n_2} \dots P_{Inp}(w_r)^{n_r}$$

$$N = \sum_i n_i = \text{summary size}$$

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?

	<i>Pyramid</i>	<i>Responsiveness</i>
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