

Automatic summarization of news

Lecture 12

Cis 530/430

Standard distinctions

- Generic vs query-focused
 - Generic summaries do not have a specific information need in mind
- Single- vs. multi-document
 - The input is either one news article, or
 - Multiple articles on the same topic (think google news)
 - Single document summarization of news is a very difficult task. The baseline of taking the beginning of the article is too difficult to beat!

Project ideas (1)

- What should a multi-document summary contain
 - Information common across many of the sources?
 - Important information that appears only in one?
 - In most existing research models (gold-standards) are created by simply asking people to write a summary. They seem to be going for the first type. But wouldn't the other kind be more useful?

How much condensation is required?

- A typical summary is about a paragraph long (100—200 words)
 - This might be too long for a single news article
 - Maybe a headline of 10-20 words?
 - Or too short, for a summary of a book or a meeting

Improving search result quality by customizing summary length [ACL-08:HLT]

query	Who was the first person to scale K2?
Paragraph	An Italian expedition finally succeeded in ascending to the summit of K2 on July 31, 1954. The expedition was led by Ardito Desio, although the two climbers who actually reached the top were Lino Lacedelli and Achille Compagnoni. The team included a Pakistani member, Colonel Muhammad Ata-ullah. He had been a part of an earlier 1953 American expedition which failed to make the summit because of a storm which killed a key climber, Art Gilkey. On the expedition also was the famous Italian climber Walter Bonatti. He proved vital to the expeditions success in that he carried vital oxygen to 26,600ft for Lacedelli and Compagnoni. His dramatic bivouac, at that altitude with the equipment, wrote another chapter in the saga of Himalayan climbing.
Sentence	The expedition was led by Ardito Desio, although the two climbers who actually reached the top were Lino Lacedelli and Achille Compagnoni.
Phrase	Lino Lacedelli and Achille Compagnoni

Answer Type	Answer Length					
	Phrase	Sentence	Paragraphs	Article	List	Combination
Person	1,362	735	570	378	419	68
Organization	153	172	295	165	432	51
Time	964	486	176	65	126	21
Number	2,075	964	362	88	158	50
GeoLocation	552	399	269	126	389	78
Place	128	121	173	87	295	33
Resource	104	136	733	273	959	256
Website	243	168	101	52	297	61
Purchase	200	318	780	295	1,231	276
Gossip	86	133	366	156	85	51
NatLang	946	479	186	21	171	26
GeneralInfo	396	861	3,197	3,244	1,075	359
Advice	43	164	1,257	1,086	357	151
ReasonCause	50	102	755	546	88	69
YesNo	392	281	306	73	12	8
Other	115	61	140	157	88	29
Unjudgable	59	47	36	18	18	556

Project ideas (2)

- Come up with a classification scheme about the type of input the system can get
 - Opinions, about a person, natural disaster, politics
- Incorporate in the system a confidence score
 - How good does the system think is the summary that it produced
- Instead of predefining summary length, let this be decided by the system.

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

Summary 1; rated poor

- P1 Thousands of people have attended a ceremony in Nairobi commemorating the first anniversary of the deadly bombings attacks against U.S. Embassies in Kenya and Tanzania.
- P2 Saudi dissident Osama bin Laden, accused of masterminding the attacks, and nine others are still at large.
- P3 President Clinton said, The intended victims of this vicious crime stood for everything that is right about our country and the world.
- P4 U.S. federal prosecutors have charged 17 people in the bombings.
- P5 Albright said that the mourning continues.
- P6 Kenyans are observing a national day of mourning in honor of the 215 people who died there.

Summary 2; rated good

- P1 Thousands of people have attended a ceremony in Nairobi commemorating the first anniversary of the deadly bombings attacks against U.S. Embassies in Kenya and Tanzania. Kenyans are observing a national day of mourning in honor of the 215 people who died there.
- P2 Saudi dissident Osama bin Laden, accused of masterminding the attacks, and nine others are still at large. U.S. federal prosecutors have charged 17 people in the bombings.
- P3 President Clinton said, “The intended victims of this vicious crime stood for everything that is right about our country and the world”. Albright said that the mourning continues.

Automatic summary edits

- Some expressions might not be appropriate in the new context
 - References:
 - he
 - Putin
 - Russian Prime Minister Vladimir Putin
 - Discourse connectives
 - However, moreover, subsequently
- Requires more sophisticated NLP techniques

Before

Turkey has been trying to form a new government since a coalition government led by **Yilmaz** collapsed last month over allegations that he rigged the sale of a bank. **Ecevit** refused even to consult with the leader of the Virtue Party during his efforts to form a government. **Ecevit** must now try to build a government. **Demirel** consulted Turkey's party leaders immediately after **Ecevit** gave up.

After

Turkey has been trying to form a new government since a coalition government led by **Prime Minister Mesut Yilmaz** collapsed last month over allegations that he rigged the sale of a bank. **Premier-designate Bulent Ecevit** refused even to consult with the leader of the Virtue Party during his efforts to form a government. **Ecevit** must now try to build a government. **President Suleyman Demirel** consulted Turkey's party leaders immediately after **Ecevit** gave up.

Sentence fusion

Sentence A Post-traumatic stress disorder (PTSD) is a psychological disorder which is classified as an anxiety disorder in the DSM-IV.

Sentence B Post-traumatic stress disorder (abbrev. PTSD) is a psychological disorder caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.

Fusion 1 Post-traumatic stress disorder (PTSD) is a psychological disorder.

Fusion 2 Post-traumatic stress disorder (PTSD) is a psychological disorder, which is classified as an anxiety disorder in the DSM-IV, caused by a mental trauma (also called psychotrauma) that can develop after exposure to a terrifying event.

Sentence compression

Orig1 The Reverse Engineer Tool is available now and is priced on a site-licensing basis, ranging from \$8,000 for a single user to \$90,000 for a multiuser project site.

Comp1 The Reverse Engineer Tool is priced from \$8,000 for a single user to \$90,000 for a multiuser project site.

Orig2 Essentially, design recovery tools read existing code and translate it into the language in which CASE is conversant — definitions and structured diagrams.

Comp2 Design recovery tools read existing code and translate it into definitions and structured diagrams.

Project ideas (3)

- All these abstractive summarization techniques can turn into very exciting (i.e. difficult) projects

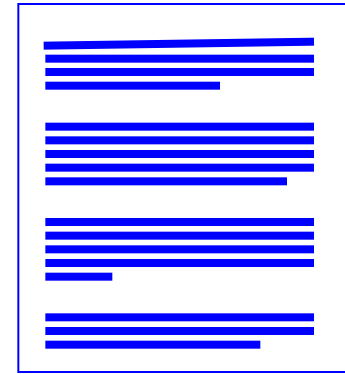
Extractive (multi-document) summarization



Input text1

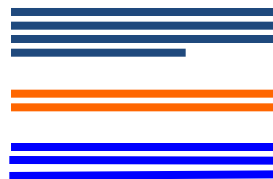


Input text2



Input text3

- 1. Selection
- 2. Ordering
- 3. Fusion



Summary

Compute Informativeness

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

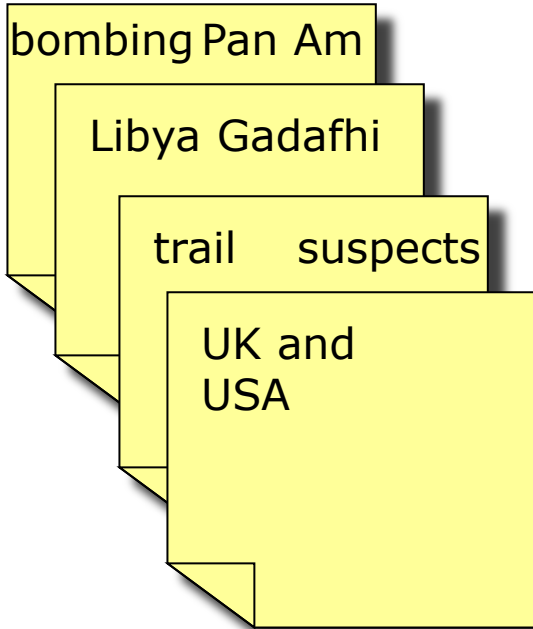
- ◆ Topic models (unsupervised)
 - Figure out what the topic of the input
 - Frequency, Lexical chains, TF*IDF
 - LSA, content models (EM, Bayesian)
 - Select informative sentences based on the topic
- Graph models (unsupervised)
 - Sentence centrality
- Supervised approaches
 - Ask people which sentences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

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

$$\textit{Weight}(\textit{Sent}) = CF(w_i \in \textit{Sent})$$

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

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

- 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

HSSSSSSSSSSSHSSSHSSHHSHHHHH

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

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

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, \dots, t_n
- Each sentence in the document is represented as a vector
$$\vec{s}_i = (w_{i1}, w_{i2}, \dots, w_{in})$$
 - If a word does not appear in the sentence, $w_{ij} = 0$
 - Else $w_{ij} = tf \cdot idf_j$
- 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	$\sum_{j=1}^k w_{1j} / k$	$\sum_{j=1}^k w_{2j} / k$		$\sum_{j=1}^k w_{nj} / k$

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

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

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

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

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

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?

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 [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

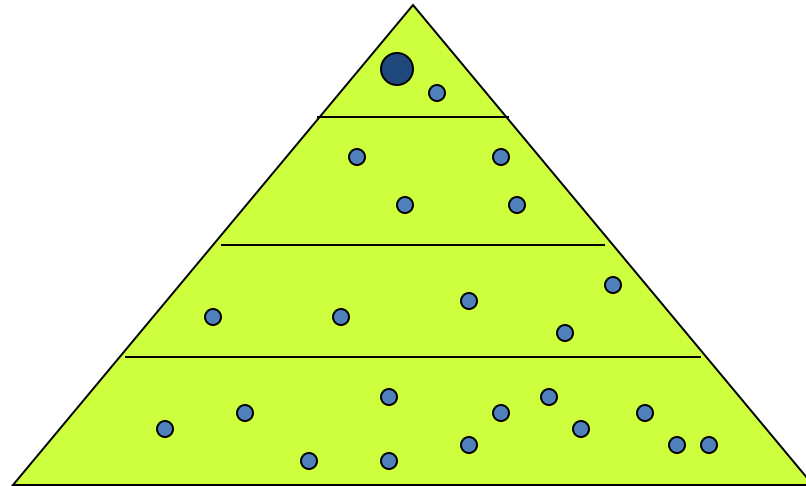
Label London was where Pinochet was arrested

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*.

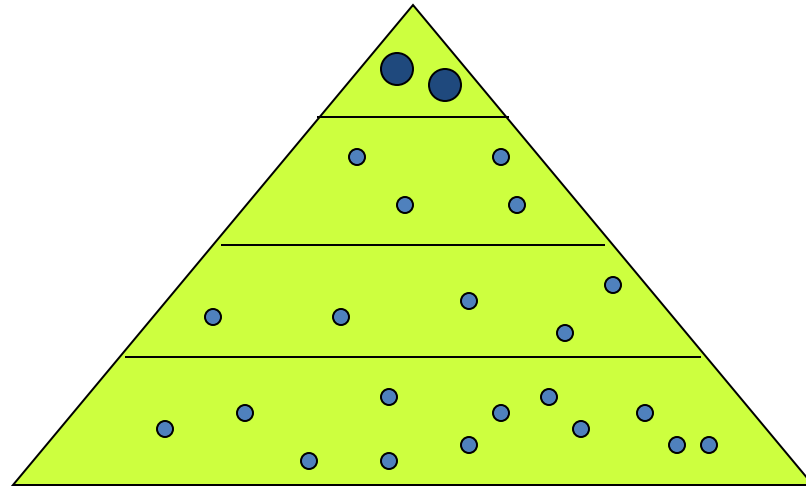
Ideally informative summary

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



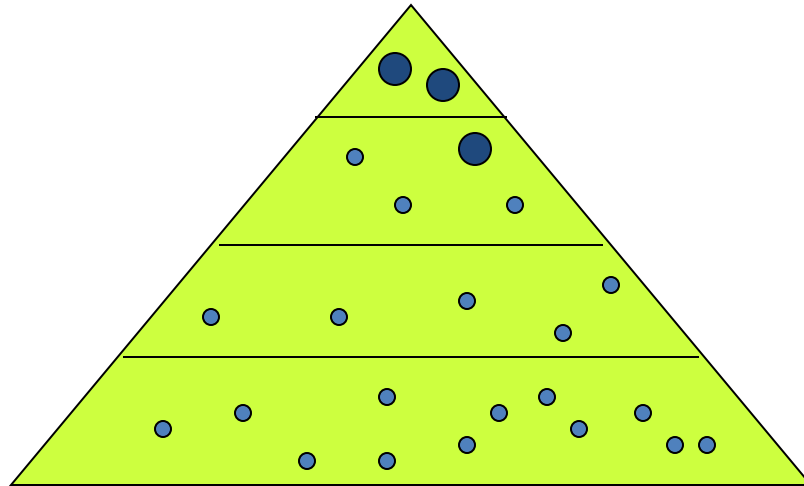
Ideally informative summary

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



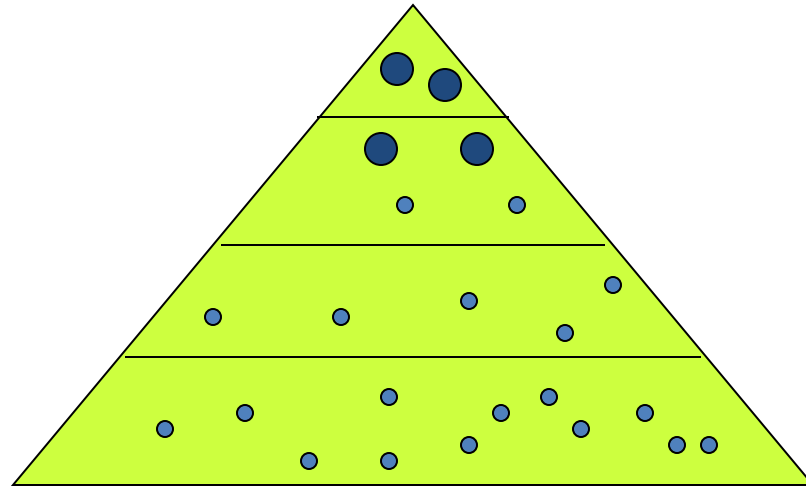
Ideally informative summary

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



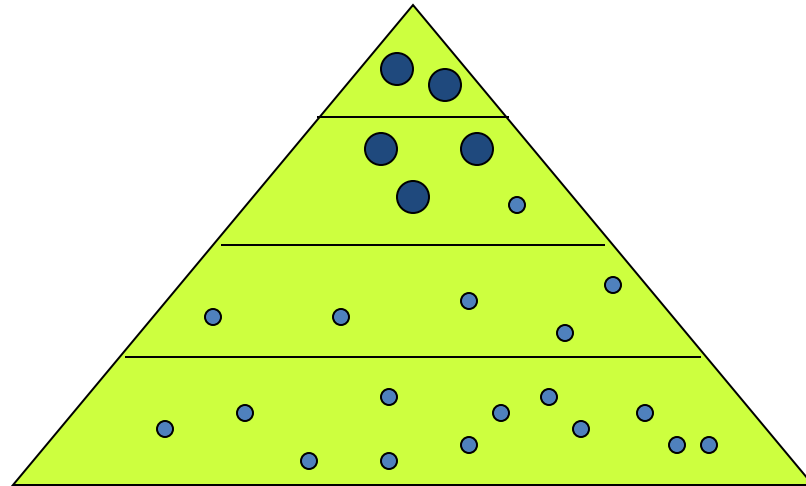
Ideally informative summary

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



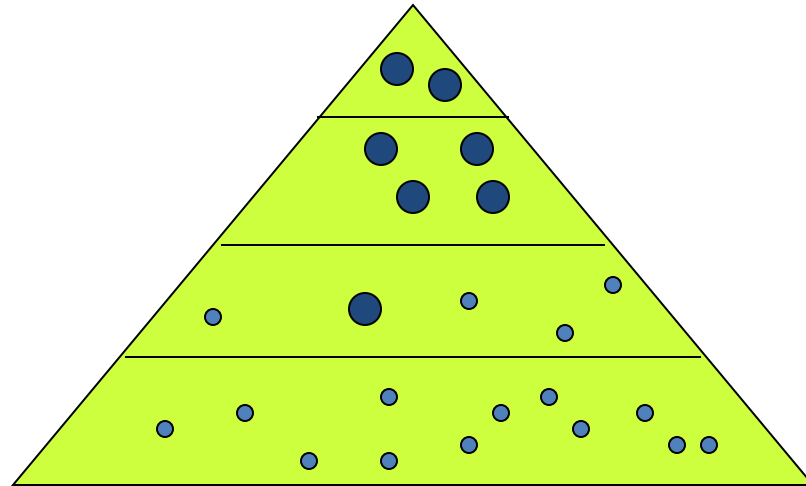
Ideally informative summary

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



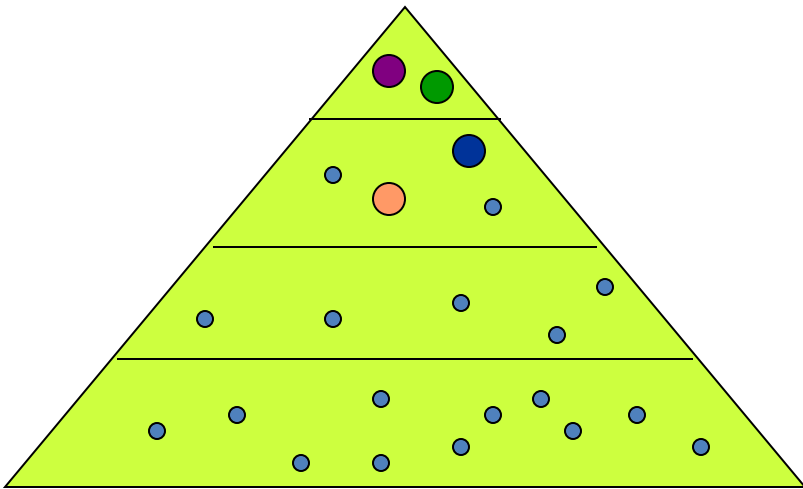
Ideally informative summary

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



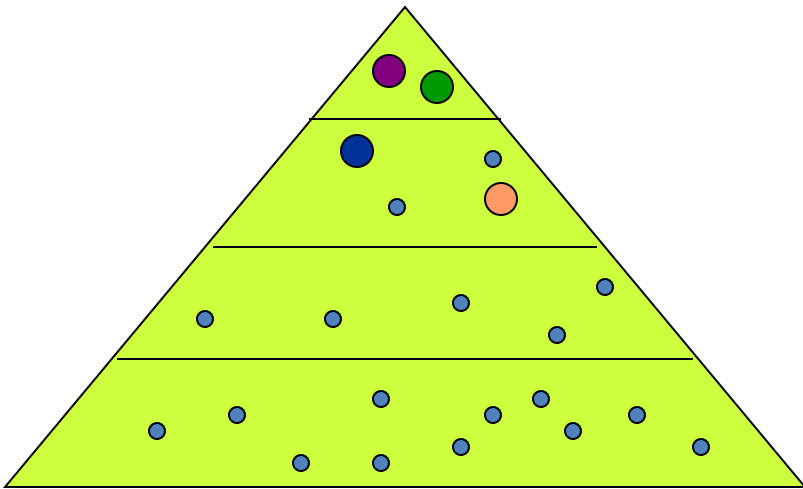
Different equally good summaries

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



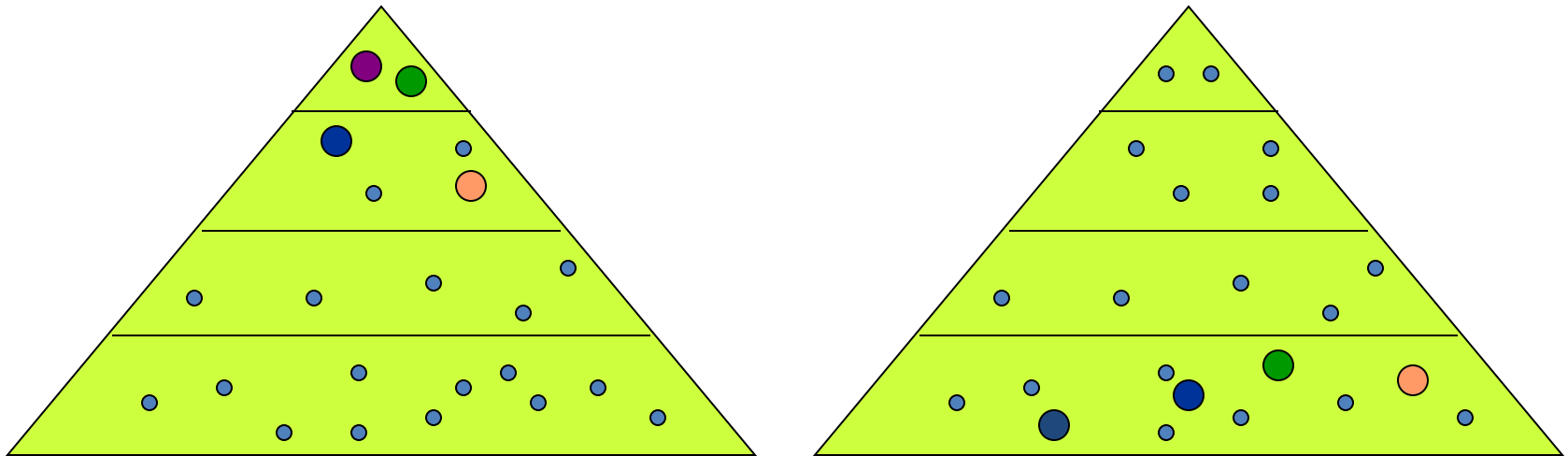
Different equally good summaries

- 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