NLP Technologies for Cognitive Computing
Lecture 2: Summarization

Devdatt Dubhashi
LAB
(Machine Learning. Algorithms, Computational Biology)
Computer Science and Engineering
Chalmers

Dealing with information overload

Adult mouse astrocytes degrade amyloid-β *in vitro* and *in*

Tony Wyss-Coray[1,2], John D. Loike[3], Thomas C. Brionne[2], Emily Lu[3], Roman Anankov[3],

Clusterin contributes to caspase-3–independent b
following neonatal hypoxia-ischemia

Byung Hee Han[1,2], Ronald B. DeMattos[1,2], Laura L. Dugan[1,2], Jeong Sook Kim-H
Robert P. Brendza[1,2], John D. Fryer[1,2], Malca Kierson[1,2], John Cirrito[1,2], Kevin Q
Judith A. K. Harmony[4], Bruce J. Aronow[5] & David M. Holtzman[1–3]

of astrocyte activation by glycolipids
NS inflammation

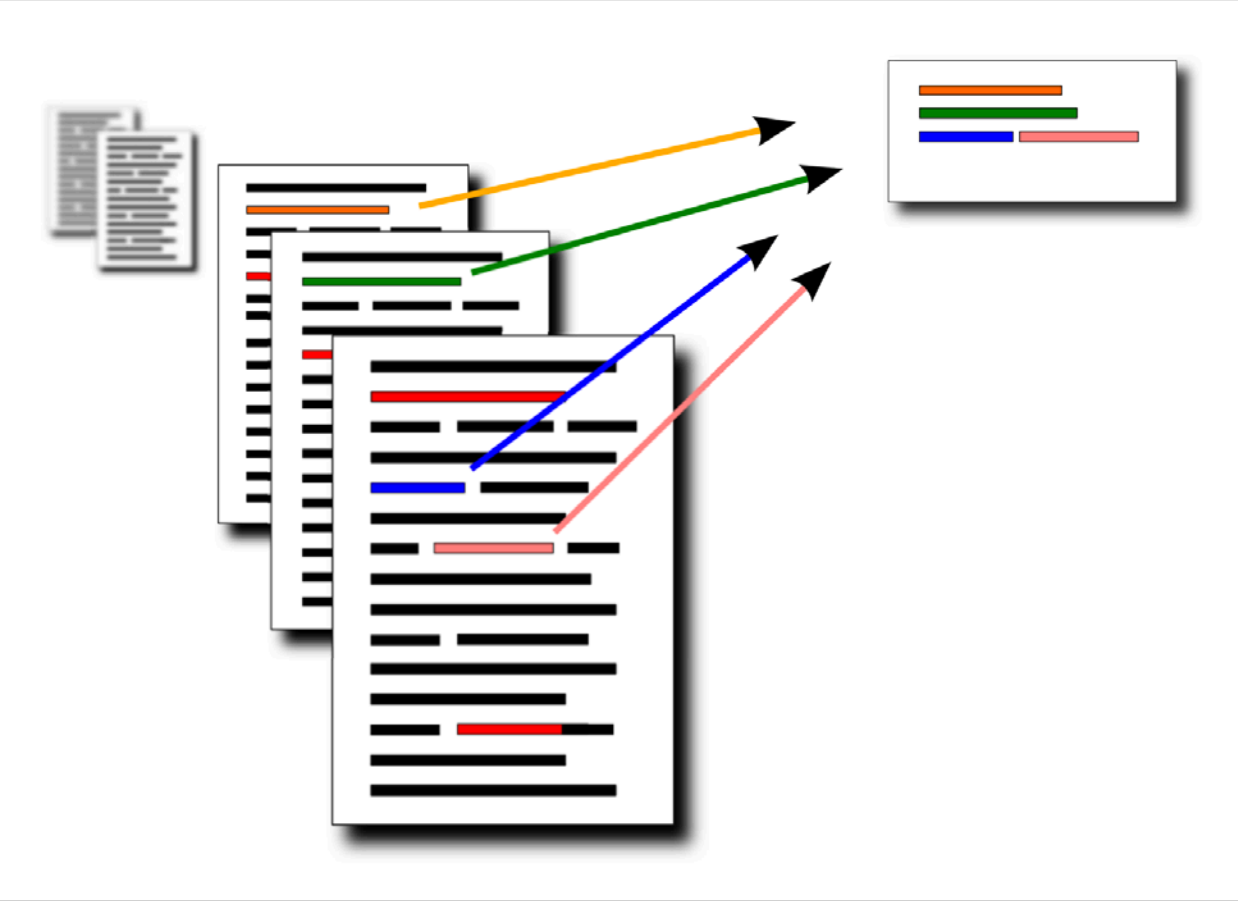
Trauger[2], Manon Blain[3], Meghan Nadeau[1], Bonny Patel[1], Jorge I Alvarez
, Ada Yeste[1], Pia Kivisäkk[1], Keith Kallas[1], Benjamin Ellezam[5], Rohit Baks
P Antel[3], Howard L Weiner[1] & Francisco J Quintana[1]

GABA from reactive astrocytes impairs memory in mouse models of Alzheimer's disease

e activation by glycolipids drives
ation

, Meghan Nadeau[1], Bonny Patel[1], Jorge I Alvarez[4],
kk[1], Keith Kallas[1], Benjamin Ellezam[5], Rohit Bakshi[1],
einer[1] & Francisco J Quintana[1]

Document summarization



M. Kågeback, O. Mogren et al,
"Extractive Summarization using Continuous Vector Space Models",
Workshop on (CVSC) EACL 2014

Olof Mogren, et ql, "Extractiv Summarization by Aggregatin Multiple similarities"
RANLP 2015

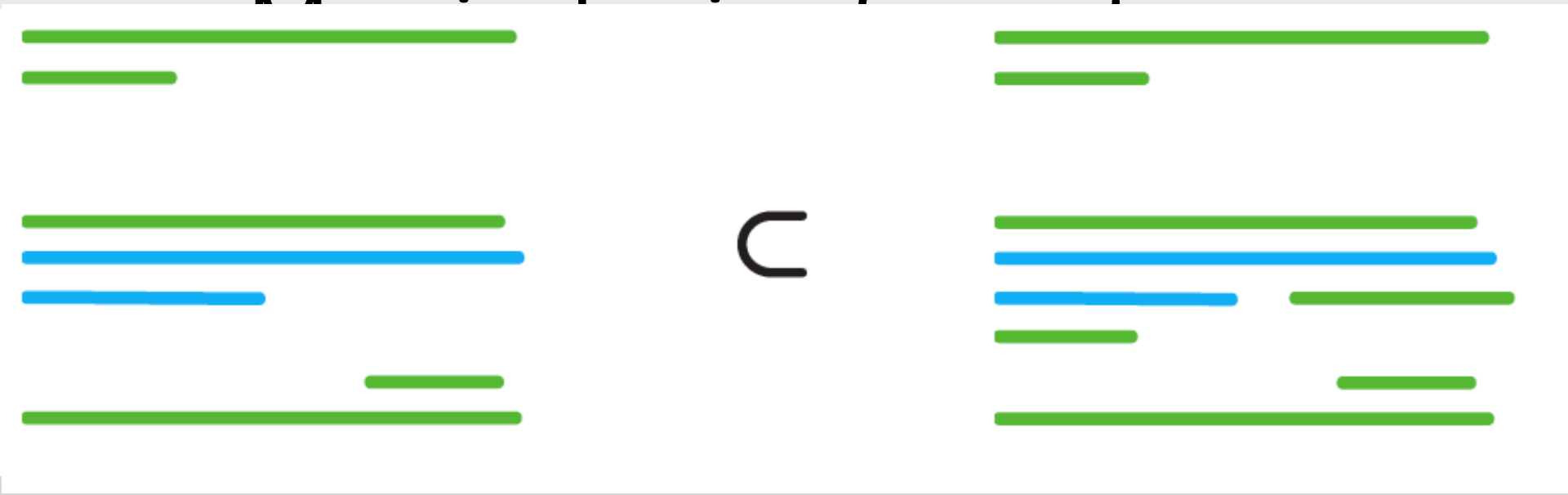
Quiz: Extractive Summarization

- If you had to pick 10 sentences to summarize a BBC report, how would you do it?
- If you had to pick sentences with a total of 100 words to form an abstract of a scientific paper?
- How is this different from usual abstracts?
- How would you evaluate a summary?

Properties of a Good Summary

- It must have high relevance
- It must be representative or diverse.

SUBMODULAR OPTIMIZATION



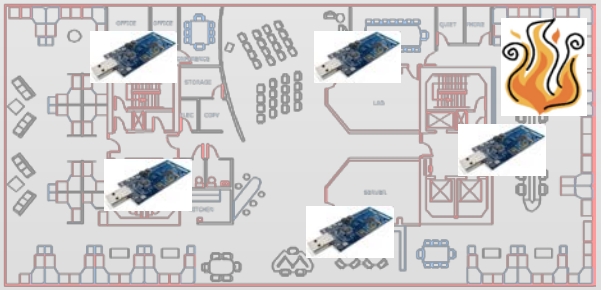
- Many natural notions of „document coverage“ are submodular [Lin & Bilmes ‘11]

Diminishing returns/submodularity

- We extract sentences (green) as a summary of the full document

- The summary on the left is a subset of the summary on the right.
- Consider adding a new (blue) sentence to each of the two summaries.
- The marginal (incremental) benefit of adding the new (blue) sentence to the smaller (left) summary is no more than the marginal benefit of adding the new sentence to the larger (right) summary.
- **diminishing returns** $\leftrightarrow$ **submodularity**

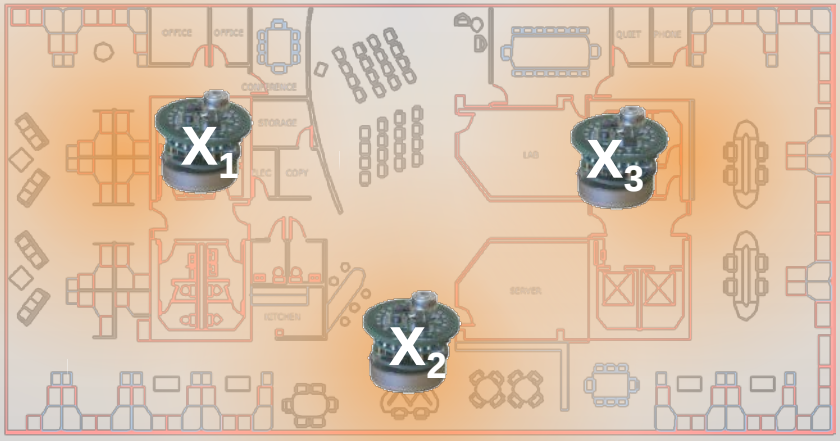
Set functions



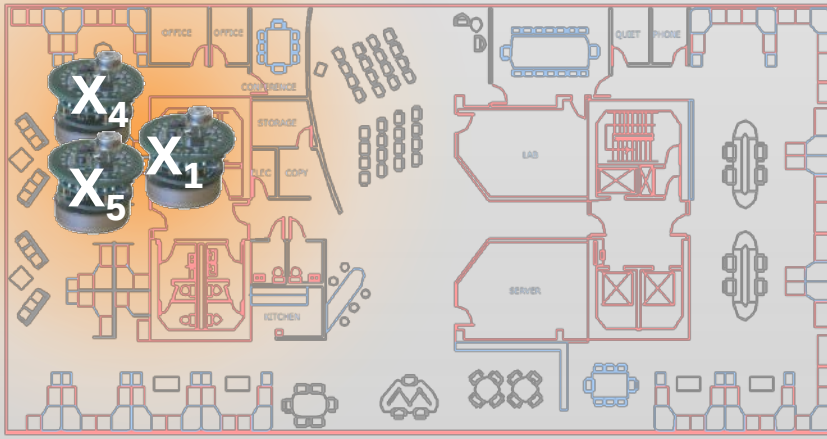
- finite ground set $V = \{1, 2, \ldots, n\}$
- set function $F : 2^V \rightarrow \mathbb{R}$
- will assume $F(\emptyset) = 0$ (w.l.o.g.)
- assume black box that can evaluate $F(A)$ for any $A \subseteq V$

Example: placing sensors

Utility $F(A)$ of having sensors at subset A of all locations



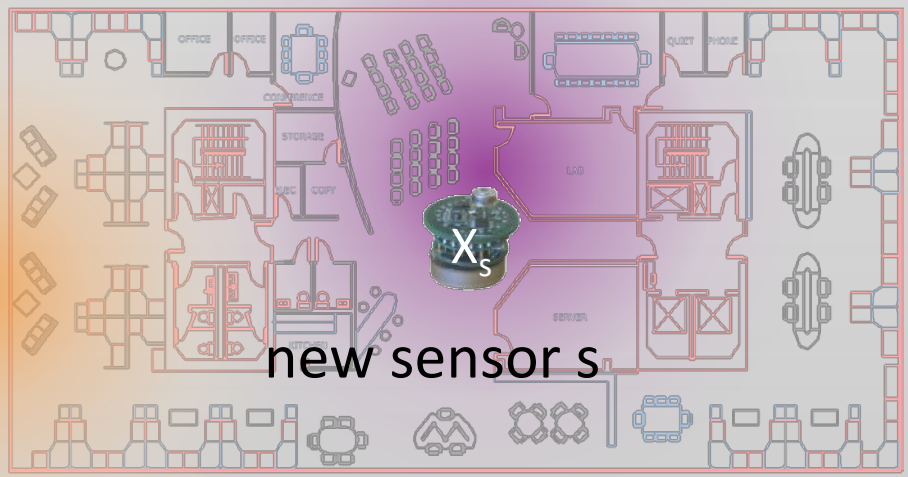
A={1,2,3}: Very informative
High value F(A)



A={1,4,5}: Redundant info
Low value F(A)

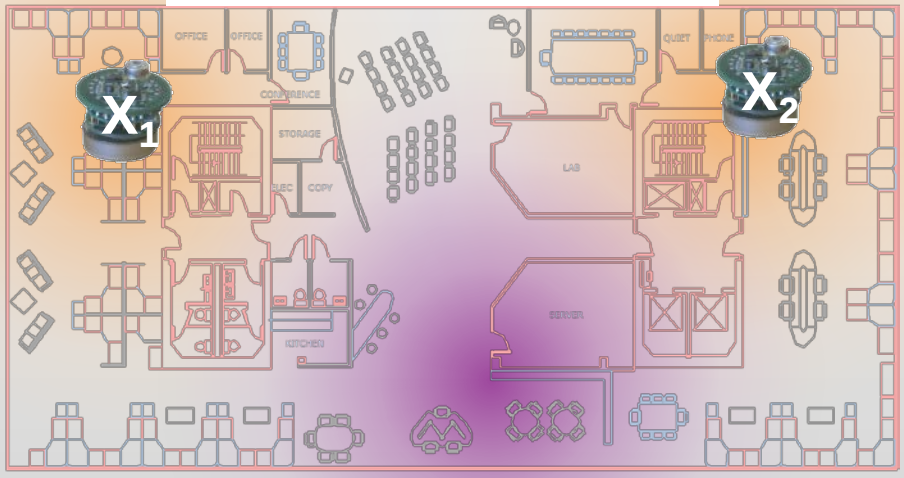
Marginal gain

- Given set function $F : 2^V \rightarrow \mathbb{R}$
- Marginal gain: $\Delta_F(s \mid A) = F(\{s\} \cup A) - F(A)$

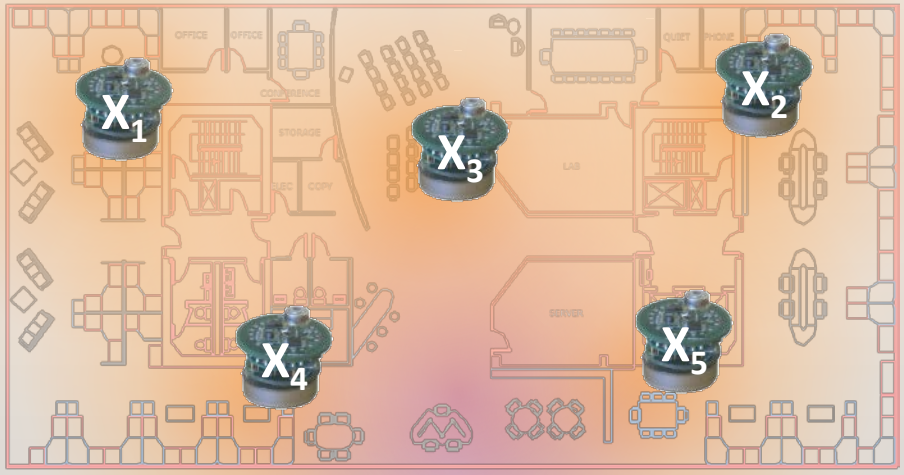


Decreasing gains: submodularity

placement A = {1,2}



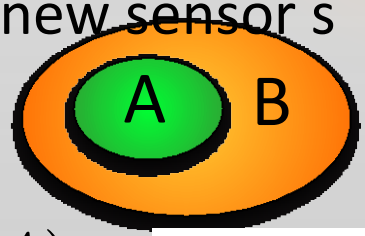
placement B = {1,…,5}



Big gain

small gain

new sensor s



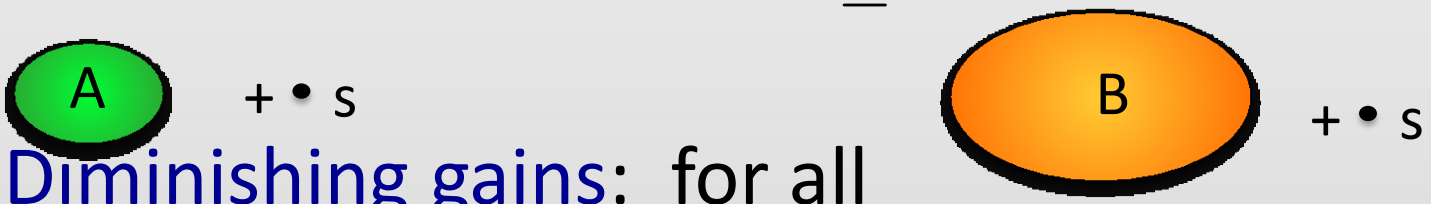
$A \subseteq B$

$$F(A \cup s) - F(A)$$

$$\Delta(s \mid A)$$

Equivalent characterizations

$$A \subseteq B$$

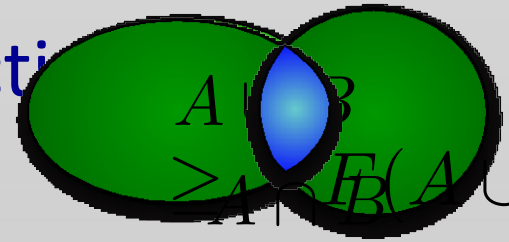


- Diminishing gains: for all

$$F(A \cup s) - F(A) \quad \geq \quad F(B \cup s) - F(B)$$

$$A, B \subseteq V$$

- Union-Intersecti



$$F(A) + F(B) \geq F(A \cup B) + F(A \cap B)$$

Submodularity

- submodularity arises in many areas: combinatorics, economics, game theory, operation research, machine learning, and (now) natural language processing.
- submodularity has many nice properties, e.g. submodularity is preserved under many natural operations and transformations (e.g.scaling, addition, convolution, etc.)

Summarization as Submodular Optimization

- Ground set *V* is the set of all sentences
- Extractive document summarization: select a small subset $S \subseteq V$ that accurately represents the entirety (ground set V).
- The summary is usually required to be length-limited.
  - c_i: cost (e.g., the number of words in sentence i),
  - b : the budget (e.g., the largest length allowed),
  - knapsack constraint: $\sum_{i \in S} c_i \leq b$
- Quality of summary: $f(S)$
- S^* = argmax $\{f(S): S \subseteq V, \sum_{i \in S} c_i \leq b\}$

Document summarization

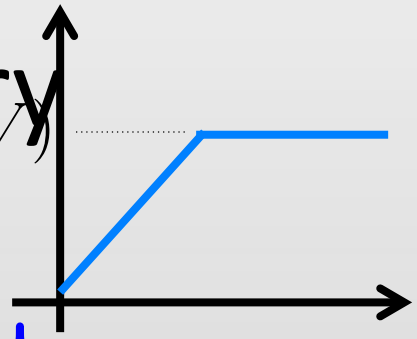
$$F(S) = R(S) + \lambda D(S)$$

Relevance

Diversity

Relevance of a summary

$$F(S) = R(S) + \lambda D(S)$$



$$R(S) = \sum_i \square C_i(S) \square$$

How well is sentence i „covered“ by S

$$C_i(S) = \sum_{j \in S} w_{i,j}$$

Similarity between i and j

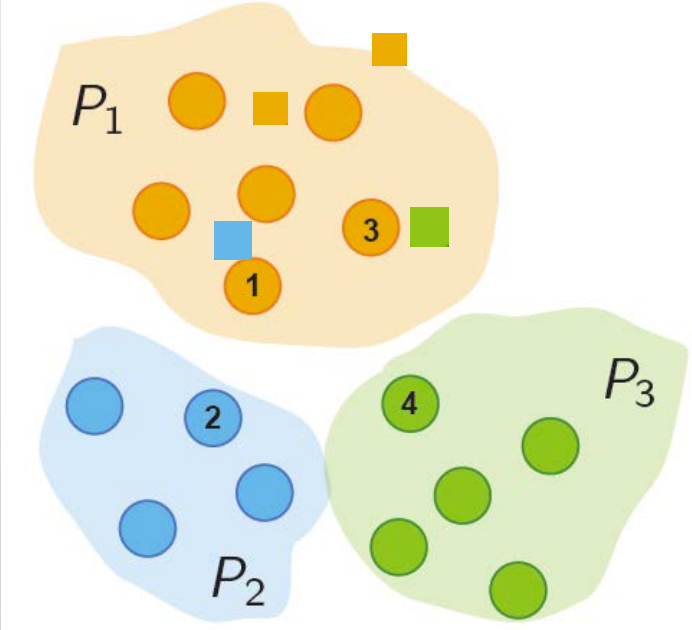
Diversity of a summary

$$D(S) = \sum_{i=1}^{K} \sqrt{\sum_{j \in P_i \cap S} r_j}$$

Relevance of sentence j to doc.

$$r_j = \frac{1}{N} \sum_{i} w_{i,j}$$

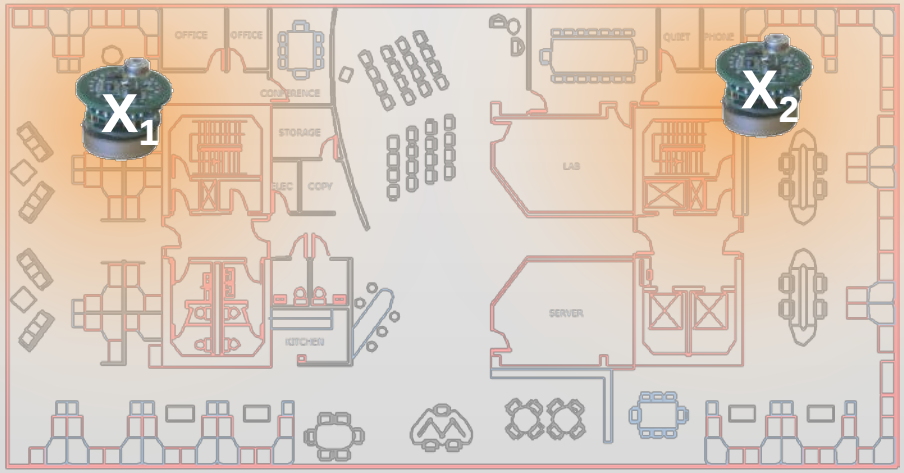
Similarity between i and j



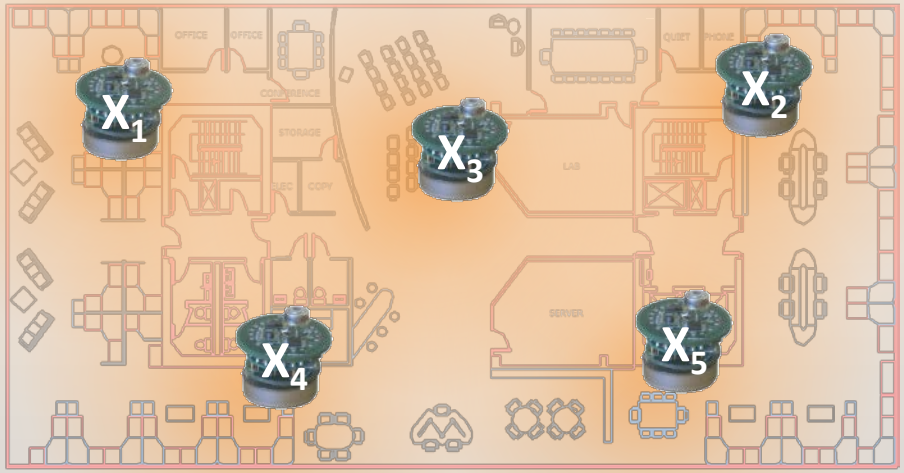
Clustering of sentences in document

Monotonicity

Placement A = {1,2}



Placement B = {1,…,5}



F is monotonic:

$$\forall A, s : \underbrace{F(A \cup \{s\}) - F(A)}_{\Delta(s \mid A)} \geq 0$$

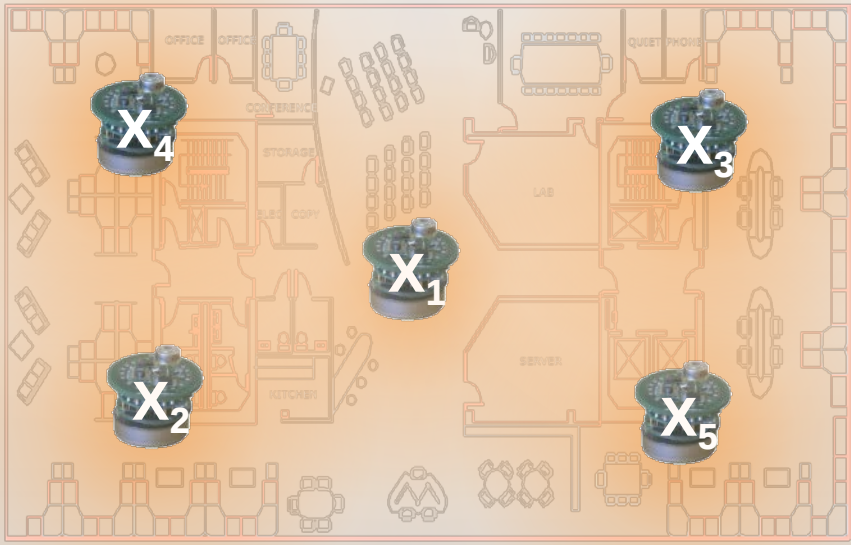
Adding sensors can only help

Cardinality constrained maximization

- **Given**: finite set V, monotone SF F
- **Want**: $\mathcal{A}^* \subseteq \mathcal{V}$ such that

NP-hard!

$$\mathcal{A}^* = \operatorname*{argmax}_{|\mathcal{A}| \leq k} F(\mathcal{A})$$

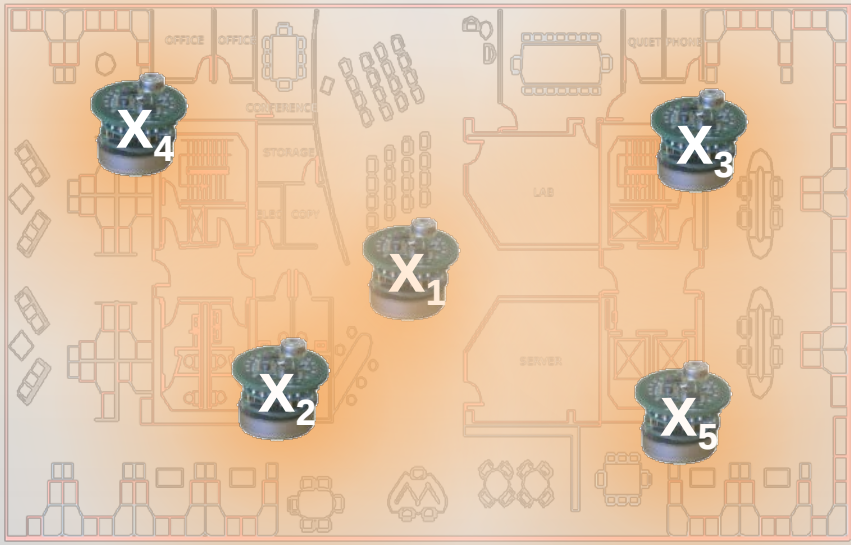


Greedy algorithm

- **Given**: finite set V, monotone SF F
- **Want**: $\mathcal{A}^* \subseteq V$ such that

$$\mathcal{A}^* = \operatorname*{argmax}_{|\mathcal{A}| \leq k} F(\mathcal{A})$$

NP-hard!



Greedy algorithm:

Start with $\mathcal{A} = \emptyset$

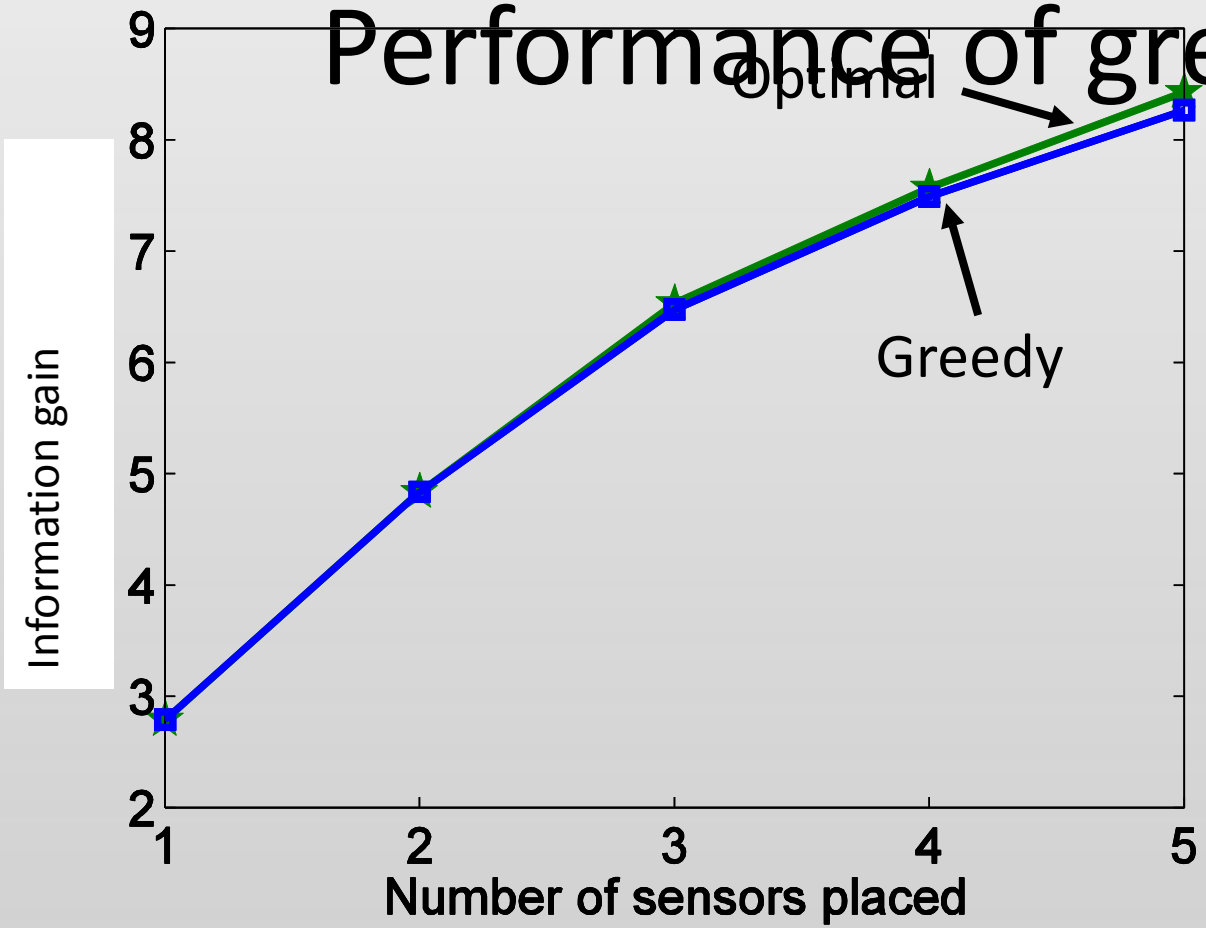
For i = 1 to k

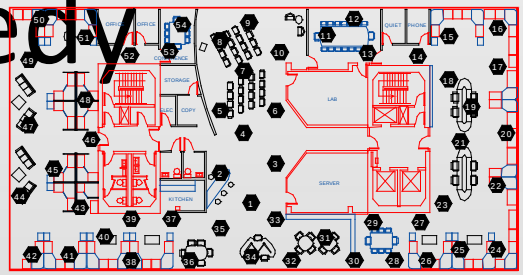
$$s^* \leftarrow \arg\max_s F(\mathcal{A} \cup \{s\})$$

$$\mathcal{A} \leftarrow \mathcal{A} \cup \{s^*\}$$

How well can this simple heuristic do?

Performance of greedy





Temperature data
from sensor network

Greedy empirically close to optimal. Why?

One reason submodularity is useful

Theorem [Nemhauser, Fisher & Wolsey '78]

For monotonic submodular functions,
Greedy algorithm gives constant factor approximation

$$F(A_{greedy}) \geq (1-1/e)\, F(A_{opt})$$

~63%

- Greedy algorithm gives near-optimal solution!
- In general, need to evaluate exponentially many sets to do better! [Nemhauser & Wolsey '78]
- Also many special cases are hard (set cover, mutual information, …)

Scaling up the greedy algorithm [Minoux ’78]

In round i+1,

- have picked $A_i = \{s_1,\ldots,s_i\}$
- pick $s_{i+1} = \text{argmax}_s\ F(A_i \cup \{s\}) - F(A_i)$

I.e., maximize “marginal benefit” $\otimes(s \mid A_i)$

$$\otimes(s \mid A_i) = F(A_i \cup \{s\}) - F(A_i)$$

Key observation: Submodularity implies

$$i \leq j \Rightarrow \otimes(s \mid A_i) \geq \otimes(s \mid A_j)$$

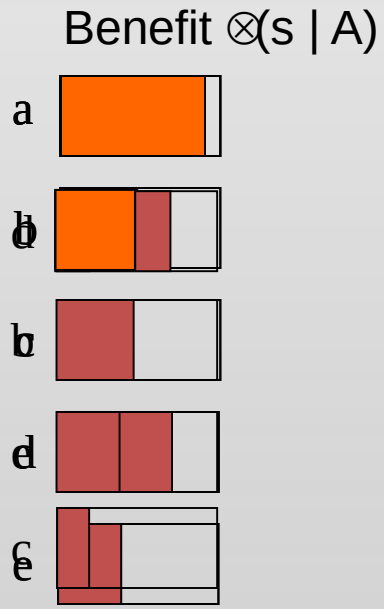
$\otimes(s \mid A_i) \geq \otimes(s \mid A_{i+1})$

Marginal benefits can never increase!

"Lazy" greedy algorithm [Minoux '78]

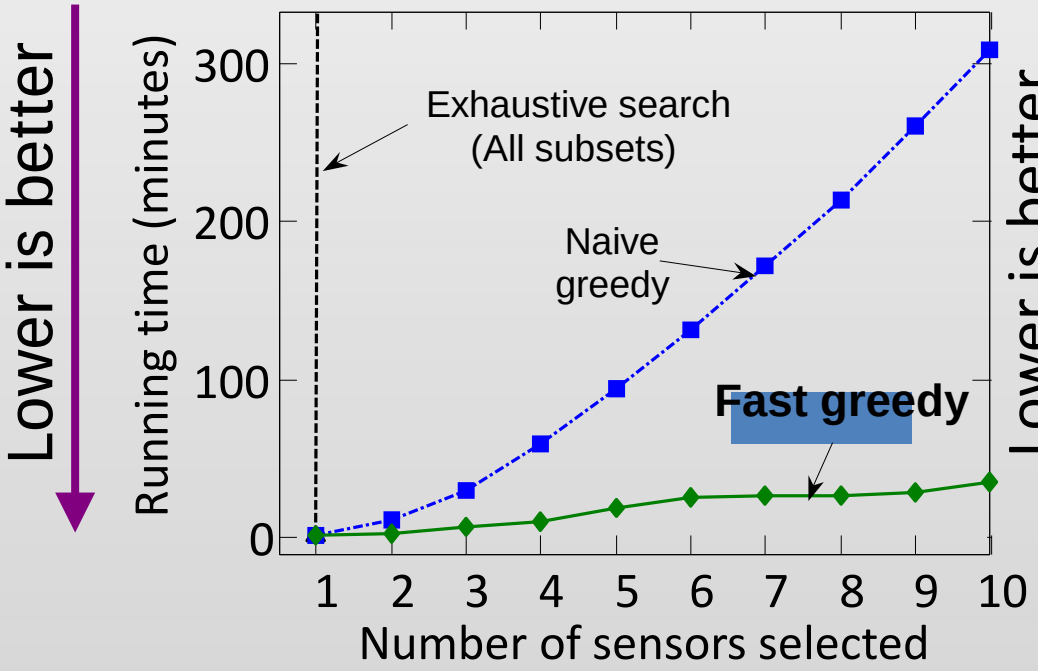
Lazy greedy algorithm:

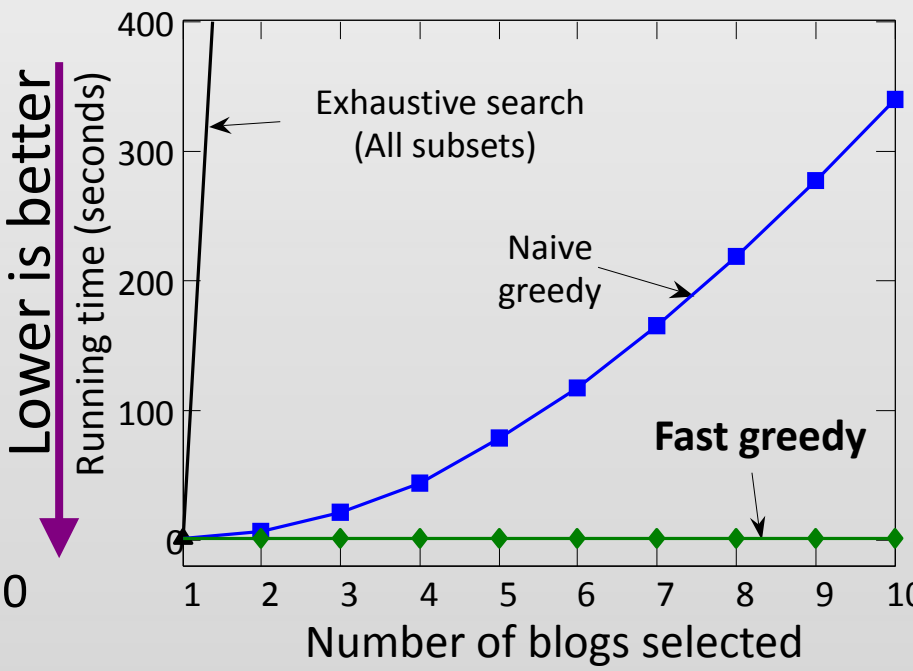
- First iteration as usual
- Keep an ordered list of marginal benefits $\otimes_i$ from previous iteration
- Re-evaluate $\otimes_i$ only for top element
- If $\otimes_i$ stays on top, use it, otherwise re-sort



Note: Very easy to compute online bounds, lazy evaluations, etc. [Leskovec, Krause et al. '07]

Empirical improvements [Leskovec, Krause et al'06]





Sensor placement

Blog selection

30x speedup

700x speedup

Evaluating Summaries: ROUGE

- ROUGE is a software package for automated evaluation of summaries (http://www.berouge.com/)
- Based co-occurrence statistics(unigram,bigram …)
- Automatic evaluation using ROUGE, between summary pairs correlates surprising well with human evaluations, based on various statistical metrics

Empirical results [Lin & Bilmes '11]

	R	F
$\mathcal{L}_1(S) + \lambda \mathcal{R}_Q(S)$	12.18	12.13
$\mathcal{L}_1(S) + \sum_{\kappa=1}^{3} \lambda_\kappa \mathcal{R}_{Q,\kappa}(S)$	**12.38**	**12.33**
Toutanova et al. (2007)	11.89	11.89
Haghighi and Vanderwende (2009)	11.80	-
Celikyilmaz and Hakkani-tür (2010)	11.40	-
Best system in DUC-07 (peer 15), using web search	**12.45**	12.29

Best F1 score on benchmark corpus DUC-07!

Can do even better using submodular structured prediction! [Lin & Bilmes '12]

COMPOSING WORD VECTORS

Similarity of Sentences

- We need a measure $w_{i,j}$ of similarity of sentences i and j.
- We have a very good measure of semantic similarity between words – cosine siimilaity of word vectors!
- How can we extend this to similarity of sentences?

Composing word vectors

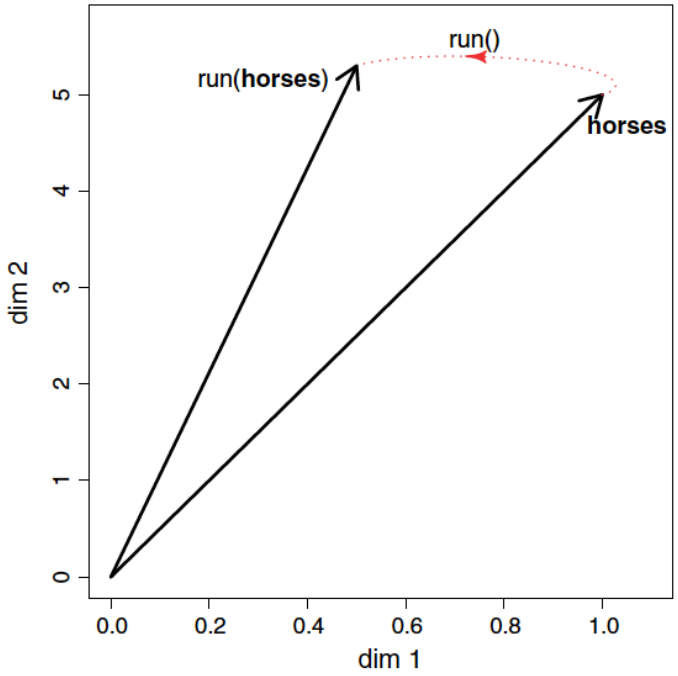


Fig 2. Left panel: composition by vector combination (in this case, addition of the *horses* and *run* vectors). Right panel: composition as function application (the verb *run* is not a vector but a function operating on vectors).

Composition using Linguistic Structures

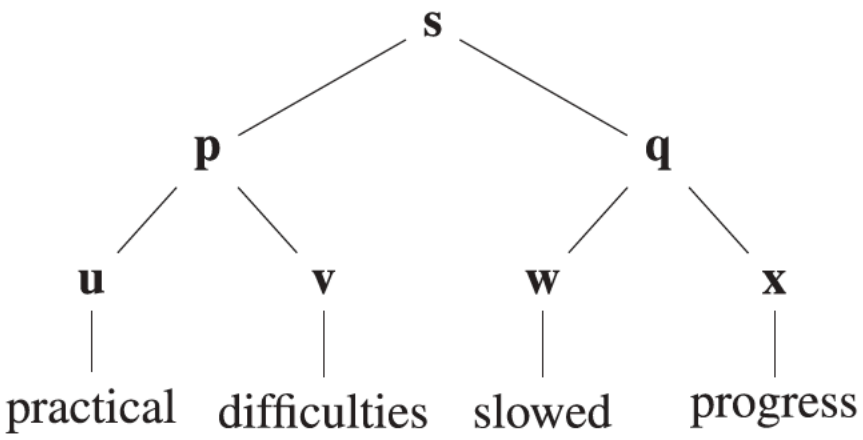
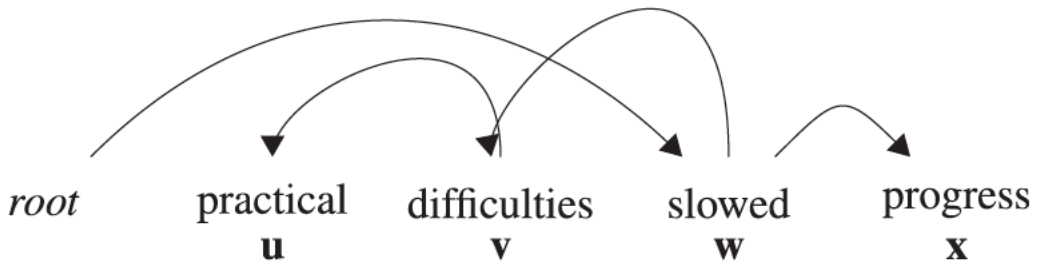


Fig. 6. Example of composition operating over parse trees.



Comparing similarity of phrases

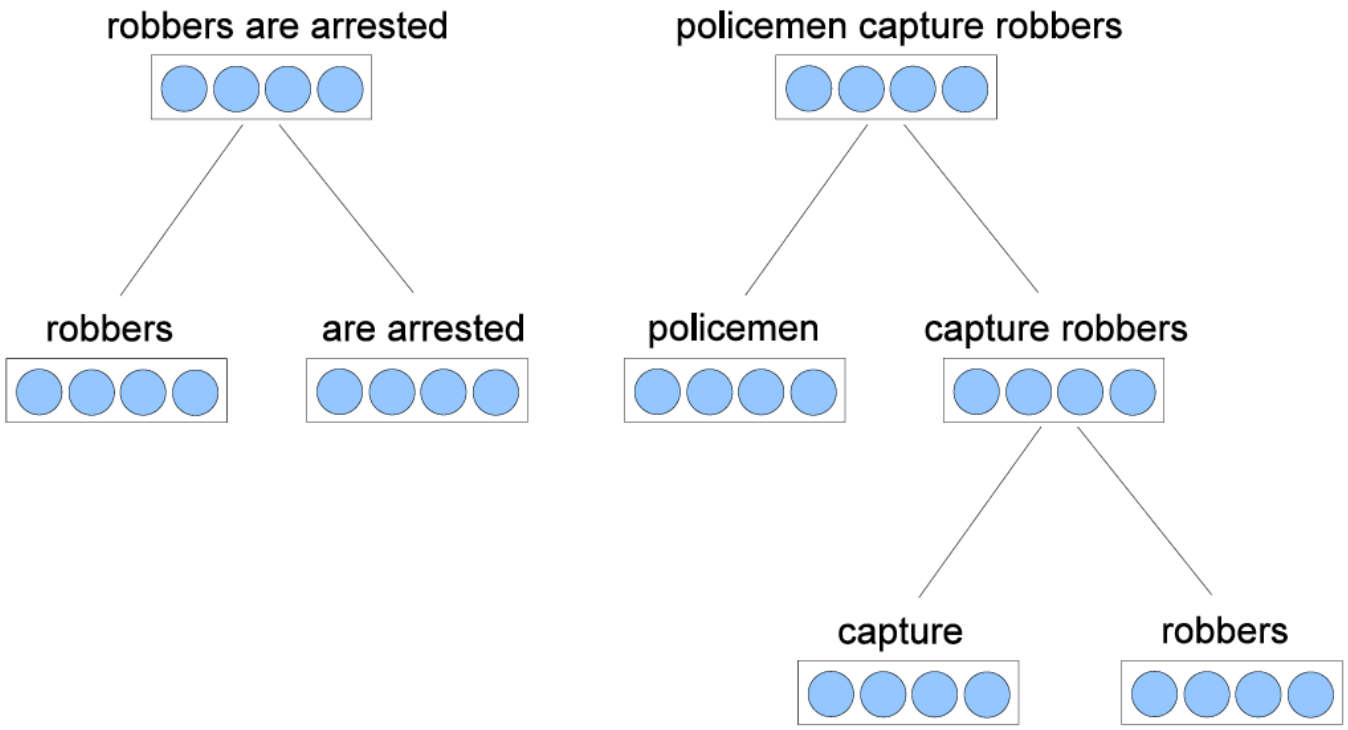
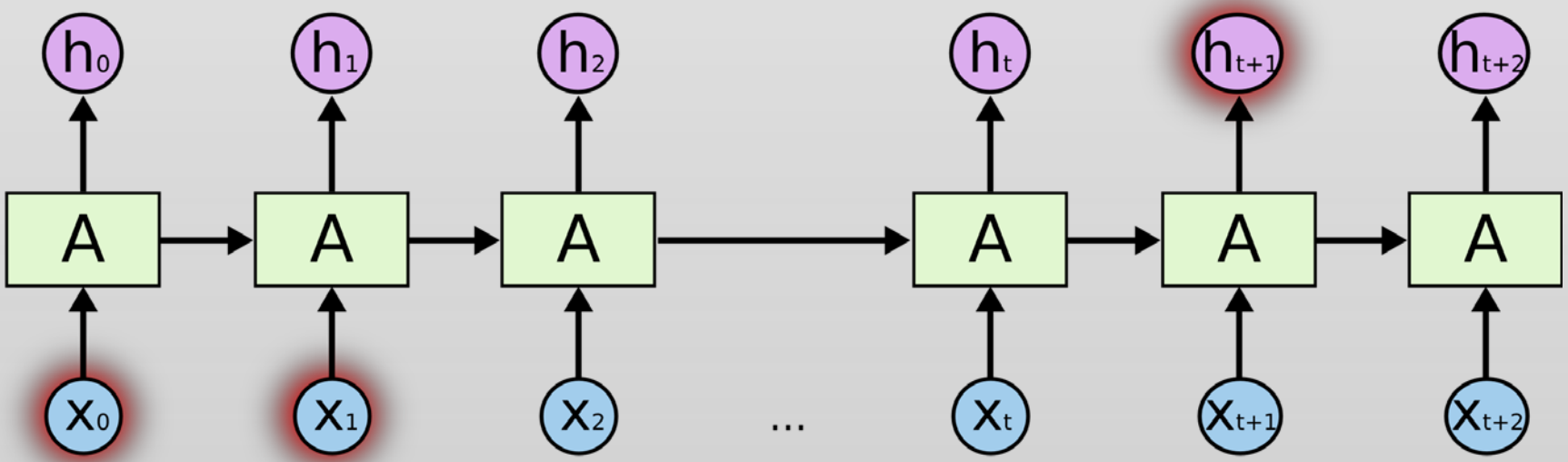


Fig 3. Distributional representations of the sentences *robbers are arrested* (left) and *policemen capture robbers* (right). Rectangles stand for vectors, possibly including those encoding functions. A more granular approach would also derive *are arrested* and the inflected forms of nouns and verbs compositionally.

Composition using LSTMs



Document Summarization

- Use submodular optimization with …
- … similarity of sentences derived by composition (in different ways) from word vector similarities.

Document Summarization

Summaries

(Approx. 40 words)

[Multiple Kernel Learning]

The report said Andreas Lubitz repeatedly set the plane for an unauthorised descent earlier that day. He had locked the flight captain out of the cockpit. Five minutes on the Duesseldorf-Barcelona flight 07:21:10 - Plane told to descend to 21,000ft

[TextRank]

The co-pilot of the Germanwings plane that

Original Text

ness Tech Science Magazine Entertainment & Arts Health Pictures World selected Africa Asia Australia Europe selec Latin America Middle East US & Canada **[Germanwings Co-pilot Lubitz 'practised rapid descent']** 21 minutes a the section Europe **[Germanwings co-pilot Andreas Lub known to have suffered depression in the past]** **[Alps plane crash]** What drives people to murder-suicide? The victims of the Germanwings plane crash Germanwings: Unanswered questions Flight 4U 9525: The final 30 minutes **[[The co-pilot of the Germanwings plane that crashed in the French Alps in March appears to have practised a rapid descent on a previous flight, a report by French investigators says.]]** **[The report said Andreas Lubitz repeatedly set the plane for an unauthorised descent earlier that day.]** Lubitz is suspected of deliberately crashing the Airbus 320, killing all 150 people on board. **[[He had locked the flight captain out of the cockpit.]]** The plane had

CHALMERS
UNIVERSITY OF TECHNOLOGY

LAB | RESEARCH GR

FINDWISE
SEARCH DRIVEN SOLUTION

http://www.cse.chalmers.se/research/databin/demonstrators.html

References

- A. Krause and D. Golovin, “Submodular Function Maximization”, in *Tractability: Practical Approaches to Hard Problems* (To appear) .
- M. Baroni, “Composition in distributional semantics”. *Language and Linguistics Compass* 7(10): 511-522.