UNIVERSITÄT DES SAARLANDES

Automatic Word Sense Discrimination

Hinrich Schütze, 1998

What is Word Sense Discrimination? How is it different from Word Sense Disambiguation? How is it calculated? How can it be evaluated? How good are the results?

Outline

- word sense disambiguation vs. discrimination
- vectorspace model with second order co-occurrence (simplest approach)
- clustering algorithms
- evaluation
- extended approaches & their evaluation
- conclusion

What is Word Sense Discrimination?

- detecting (inducing) different senses of ambiguous words
- NO sense provided/used
- particular approach here: context-group discrimination
  - finding clusters of similar word usage contexts
  - unsupervised
  - input: text, ambiguous word (+meta-parameters)
  - output: clusters of similar word usage contexts

Why could it work? - Idea

Strong Contextual Hypothesis:

Two words are semantically similar to the extent that their contextual representations are similar.

extended by

Contextual Hypothesis for Senses:

Two occurrences of an ambiguous word belong to the same sense to the extent that their contextual representations are similar.

Applications

- some require sense-*labels* like
  - correct translation of ambiguous words
  - correct pronunciation of ambiguous words
- some don't
  - Information Retrieval (selection of sense cluster in the result)

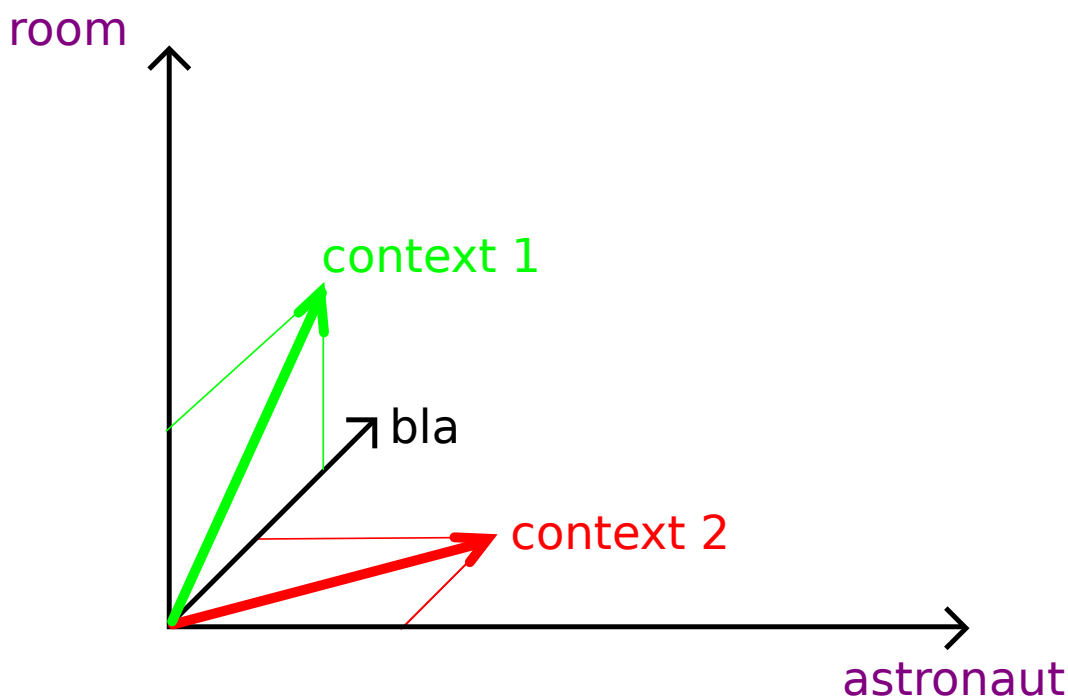
Vectorspace Model

regular first-order co-occurrence (NOT used):

... Bla | bla. A tiny room has no space for thousands of unnecessary things. Bla | bla bla ...

... Bla bla bla. | An astronaut on a mission in space sends pictures back to Earth. Bla | bla blub ...

here context radius = 6 words (Schütze uses 25)



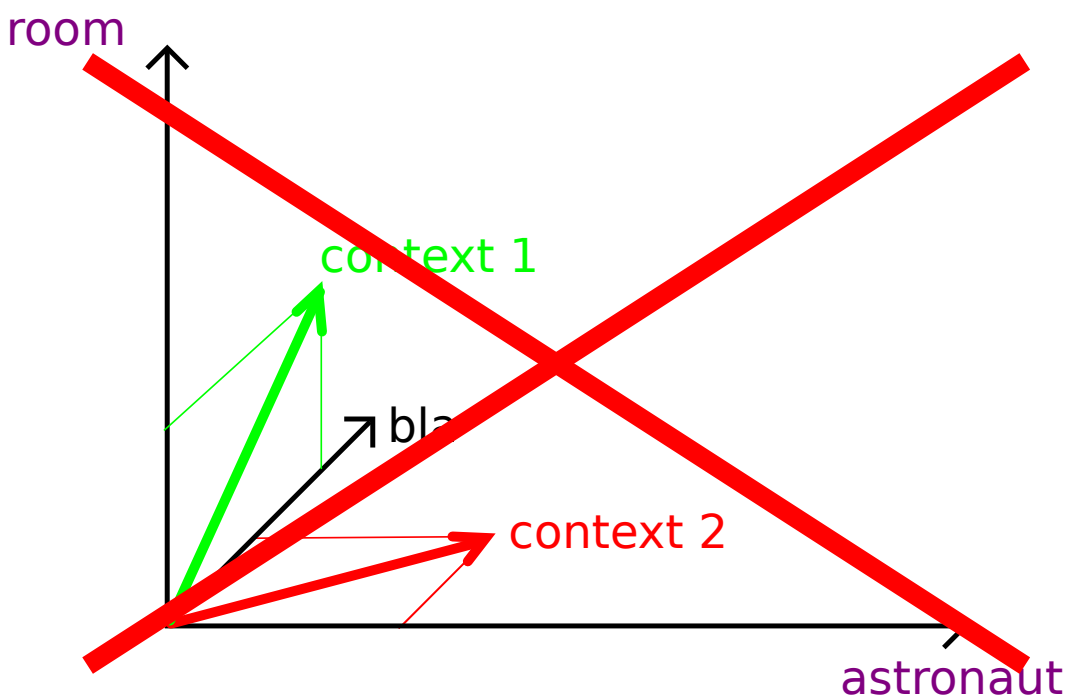
Vectorspace Model

regular first-order co-occurrence (NOT used):

... Bla | bla. A tiny room has no space for thousands of unnecessary things. Bla | bla bla ...

... Bla bla bla. | An astronaut on a mission in space sends pictures back to Earth. Bla | bla blub ...

here context radius = 6 words (Schütze uses 25)



Automatic Word Sense Discrimination

Second-order Co-occurrence

$$cos(\vec{x}, \vec{y}) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

... Bla bla bla. | An astronaut on a mission in space sends pictures back to Earth. Bla | bla blub ...

... Bla bla bla. Vikings | 1 and 2 were the first space probes successfully landing on Mars. Bla | bla blub ...

1 x astronaut
1 x mission
1 x earth

first-order co-occurrence:
no connection

1 x probes
1 x landing
1 x mars

. . .
1 x earth
. . .

1 x earth
1 x venus
1 x mars

... Venus and Mars are fellow planets of the Earth ...

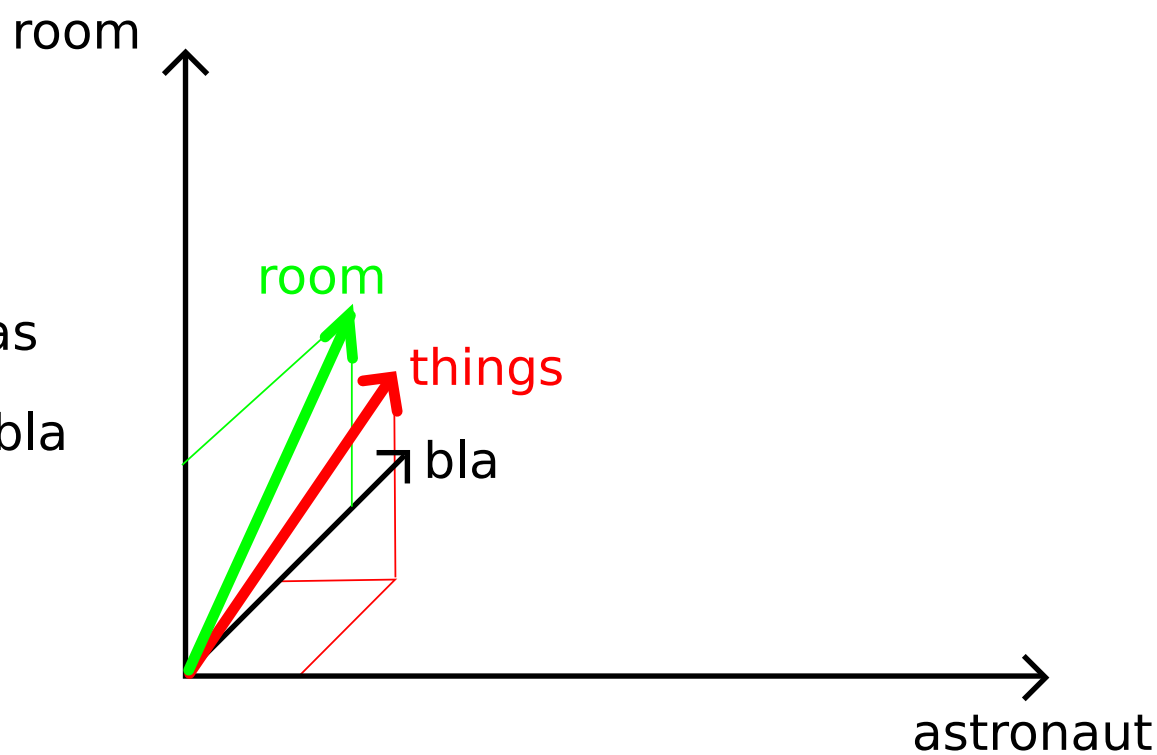
second-order co-occurrence: related

1 x mars
1 x venus
1 x earth

1 x mars

Word Vectors

... Bla | bla. A tiny room has no space for thousands of unnecessary things. Bla | bla bla ...



In the first experiments, for every ambiguous word, the 1,000 most frequent words in its contexts are used both for dimensions and word-vectors (local feature selection).

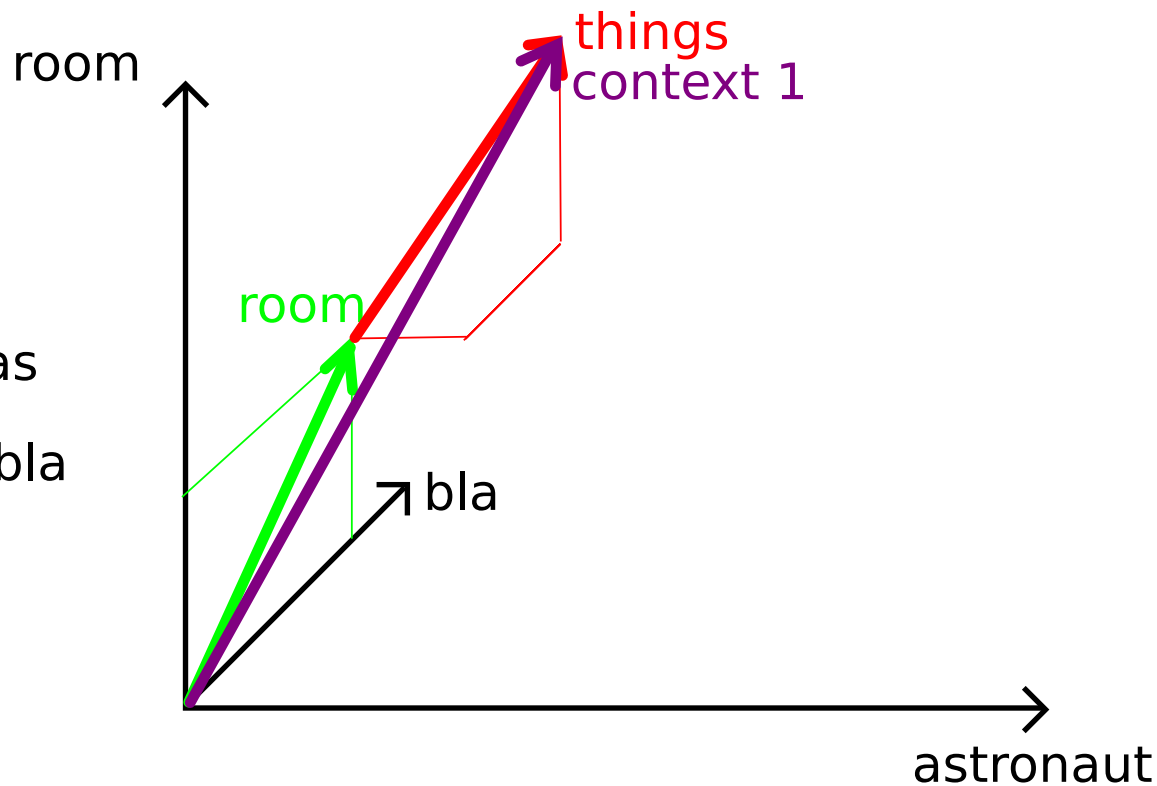
This is s 1,000 x 1,000 (first-order) co-occurrence matrix.

Term weighting by idf factor $log(\frac{\text{nr of documents}}{\text{nr of documents containing word}})$

Automatic Word Sense Discrimination

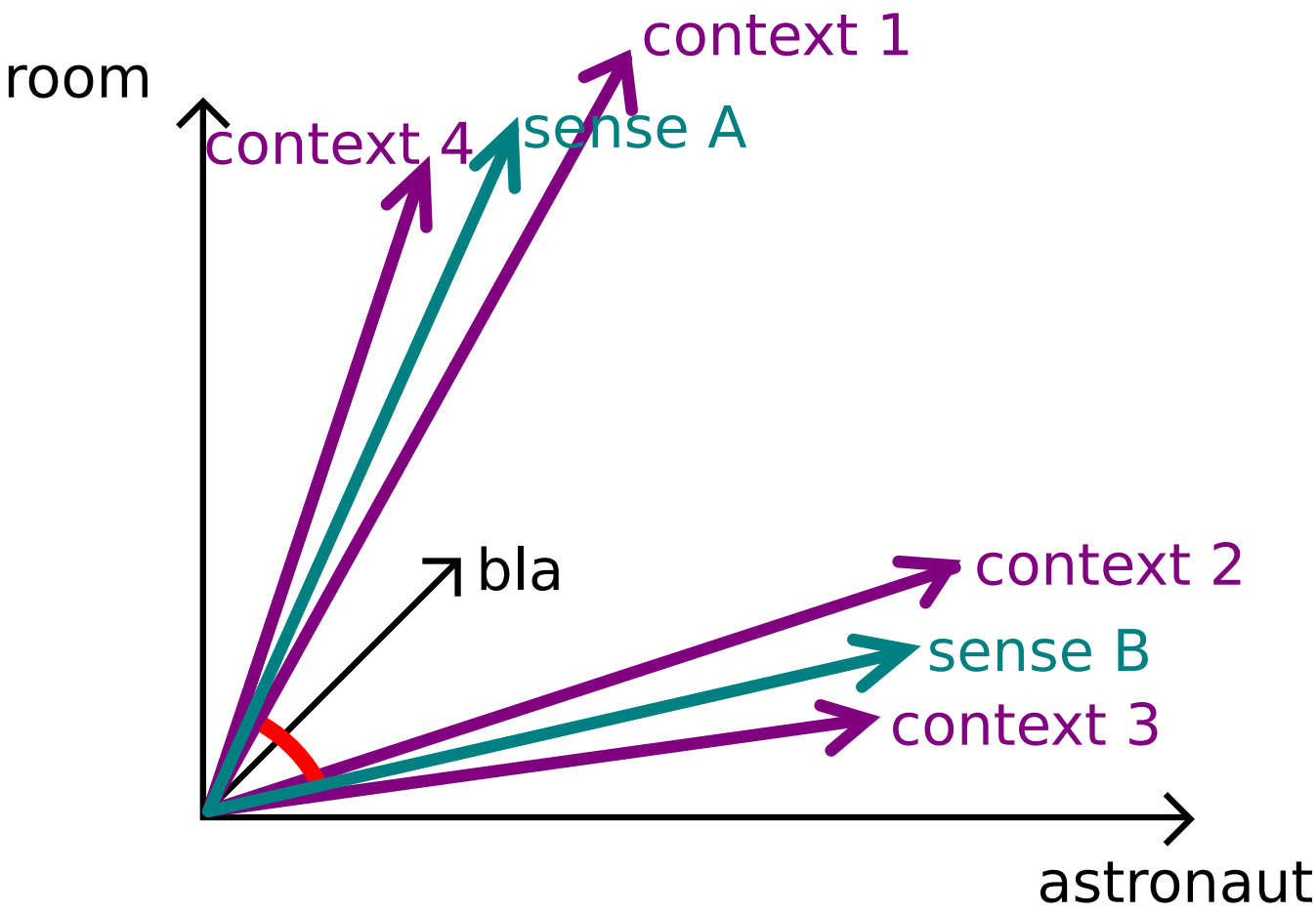
Context Vectors

... Bla | bla. A tiny room has no space for thousands of unnecessary things. Bla | bla bla ...



Context vectors are weighted sums of word vectors. This corresponds to vectors based on second-order co-occurrence.

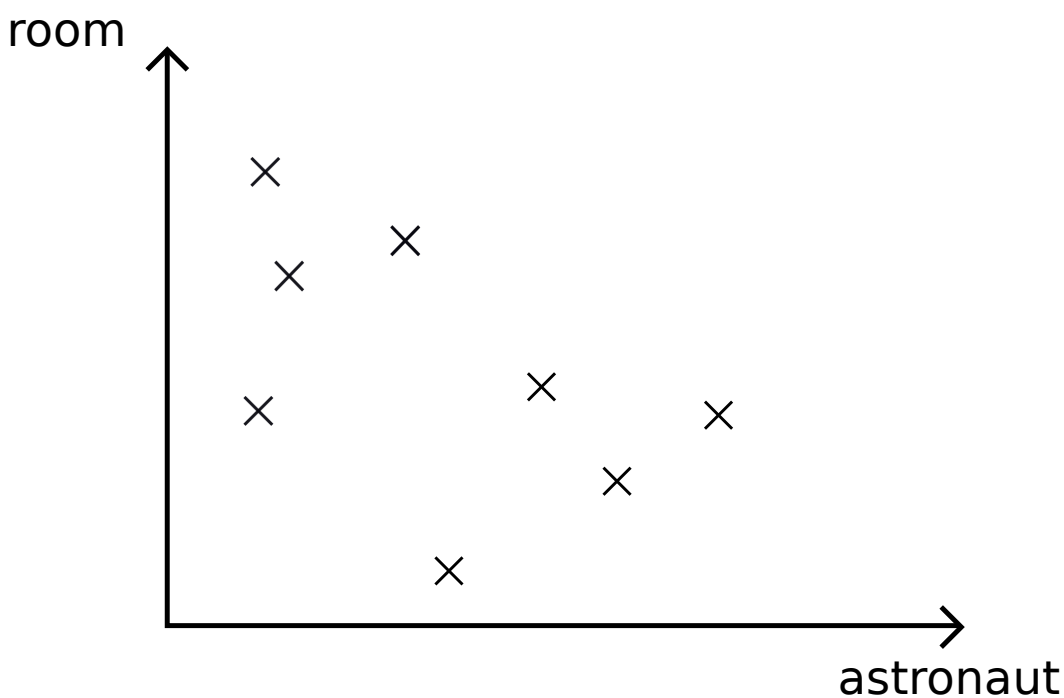
Sense Vectors



Context vectors in training data are clustered. Cluster centroids are interpreted as sense vectors.

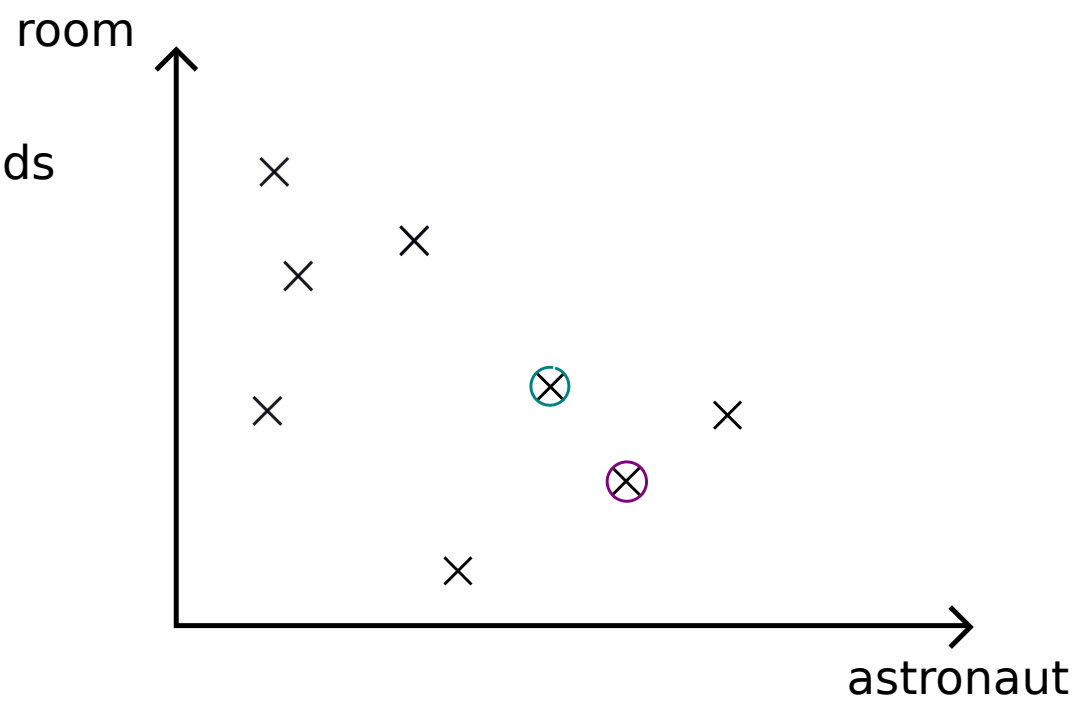
Clustering: k-Means 1/5

- unlabeled context vectors



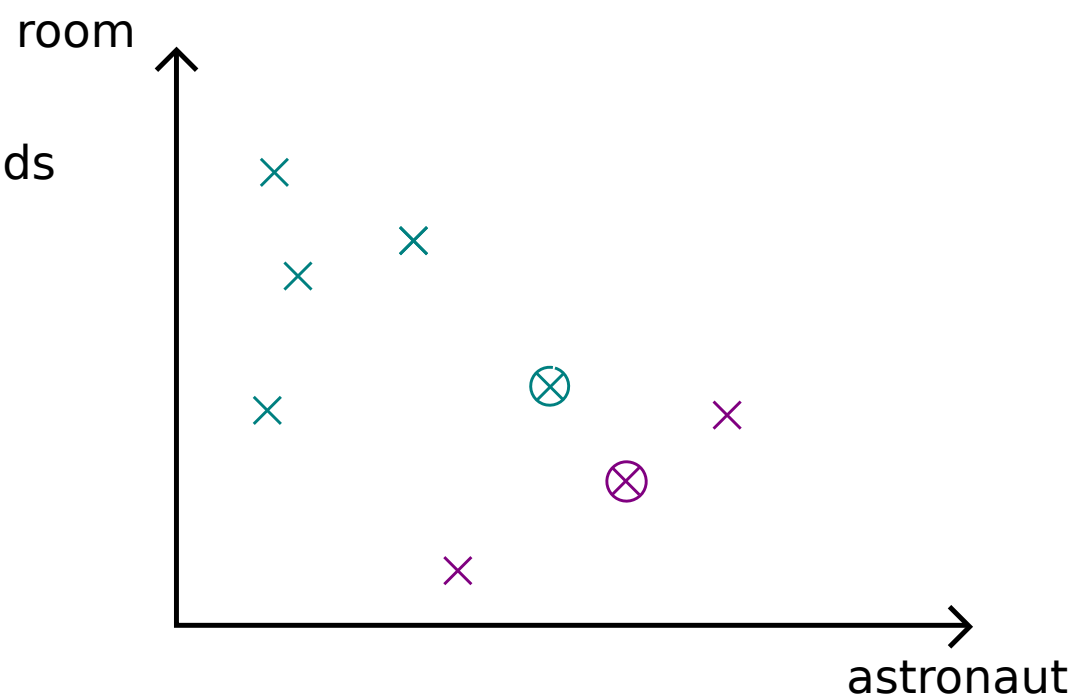
Clustering: k-Means 2/5

- unlabeled context vectors
- k random initial cluster-centroids



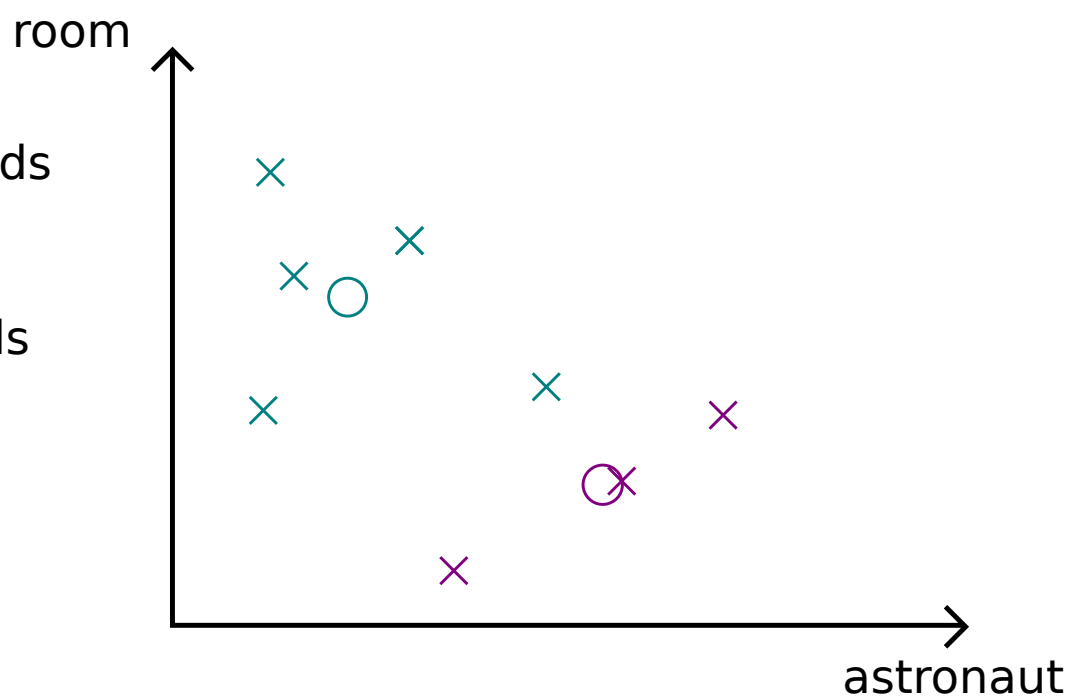
Clustering: k-Means 3/5

- unlabeled context vectors
- k random initial cluster-centroids
- E-Step: assign data points to closest centroid



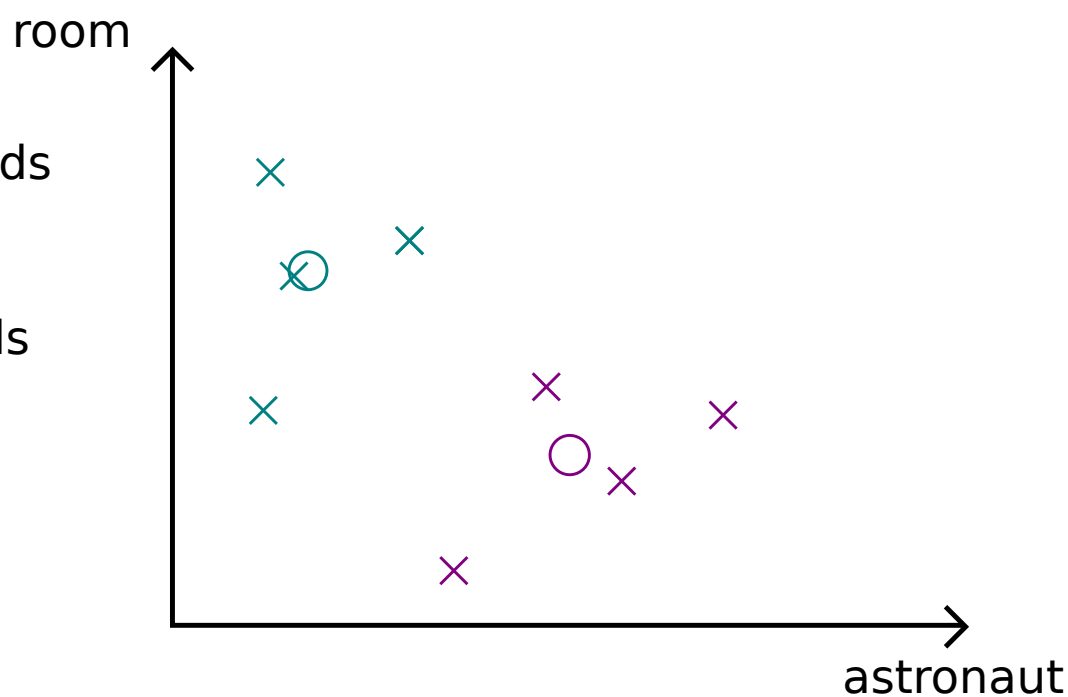
Clustering: k-Means 4/5

- unlabeled context vectors
- k random initial cluster-centroids
- E-Step: assign data points to closest centroid
- M-Step: calculate new centroids



Clustering: k-Means 5/5

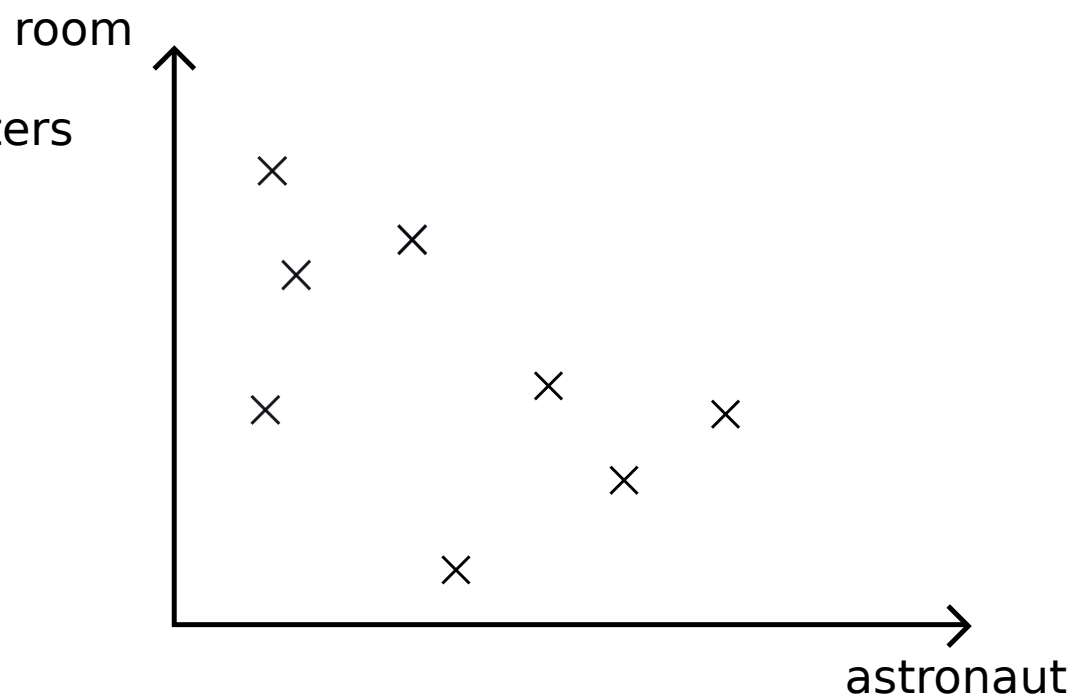
- unlabeled context vectors
- k random initial cluster-centroids
- E-Step: assign data points to closest centroid
- M-Step: calculate new centroids
- repeat 5 times
- problem: local optima
- preclustering with GAAC



Pre-Clustering:

Group Average Agglomerative Clustering (GAAC)
1/7

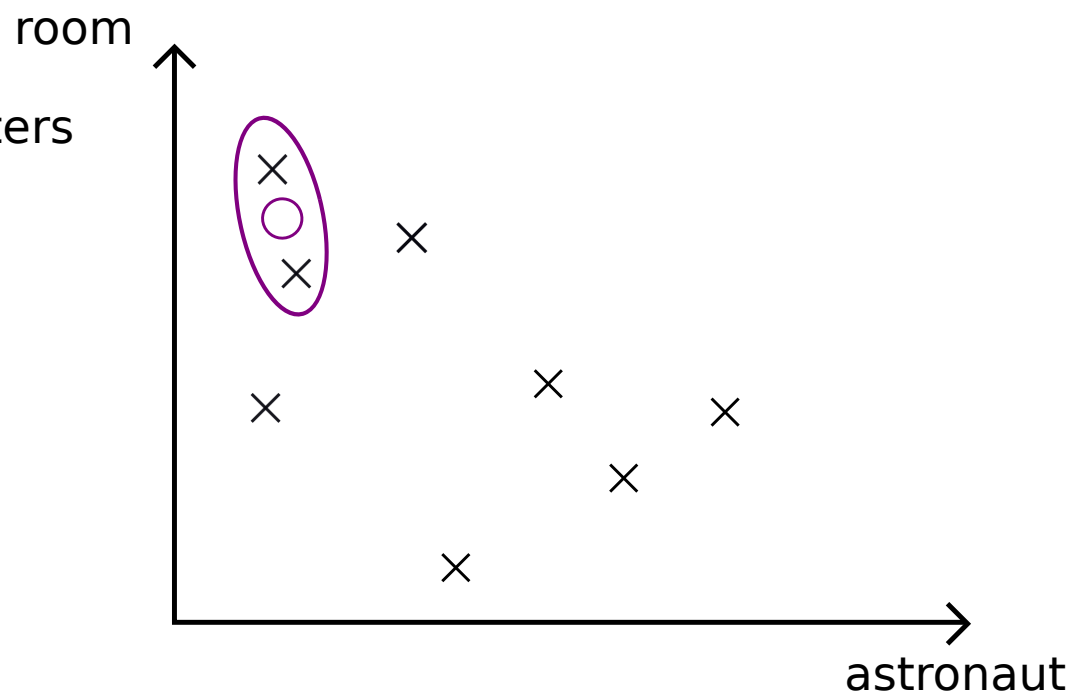
- unlabeled context vectors
- initially all datapoints are clusters



Pre-Clustering:

Group Average Agglomerative Clustering (GAAC)
2/7

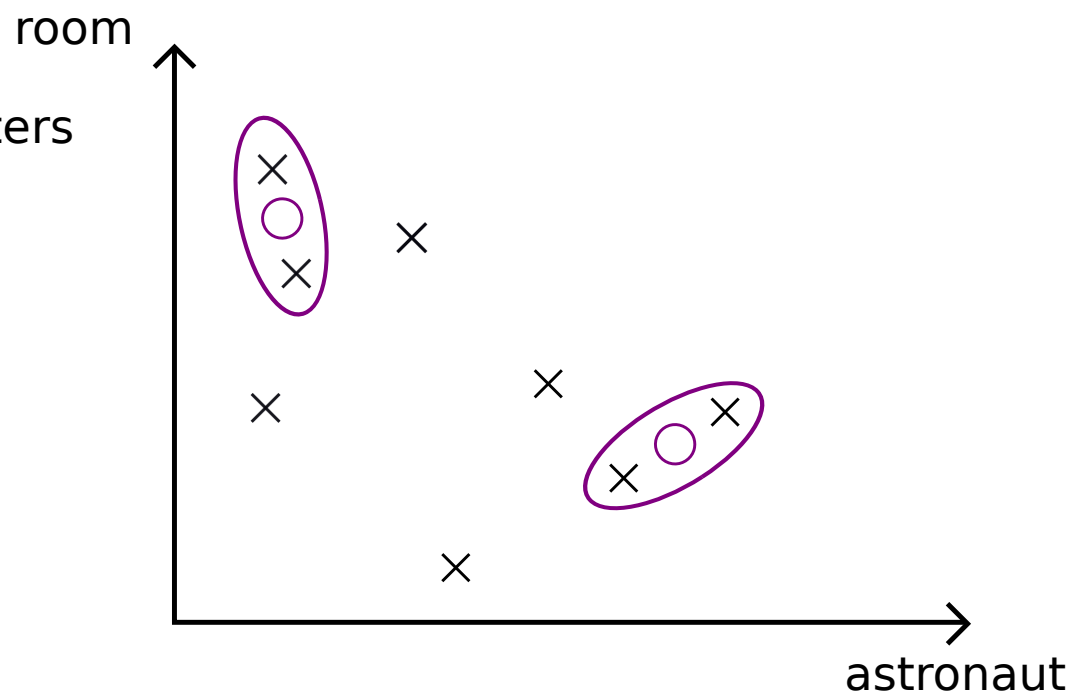
- unlabeled context vectors
- initially all datapoints are clusters
- repeatedly merge closest cluster pair (closest average distance = closest averages = closest centroids)



Pre-Clustering:

Group Average Agglomerative Clustering (GAAC)
3/7

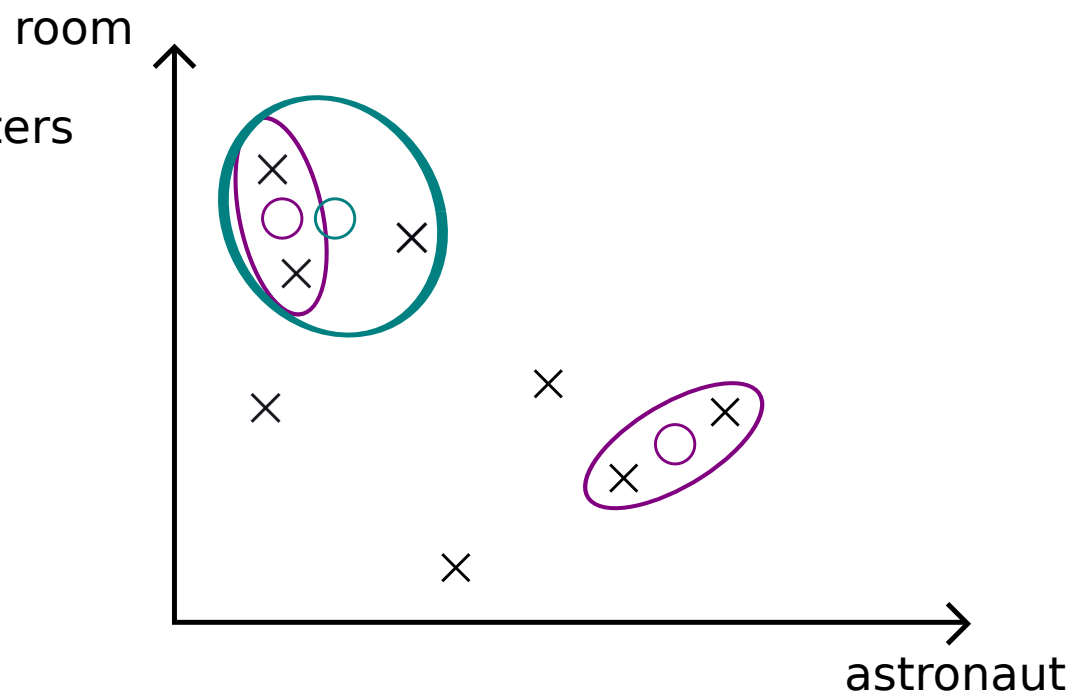
- unlabeled context vectors
- initially all datapoints are clusters
- repeatedly merge closest cluster pair (closest average distance = closest averages = closest centroids)



Pre-Clustering:

Group Average Agglomerative Clustering (GAAC)
4/7

- unlabeled context vectors
- initially all datapoints are clusters
- repeatedly merge closest cluster pair (closest average distance = closest averages = closest centroids)

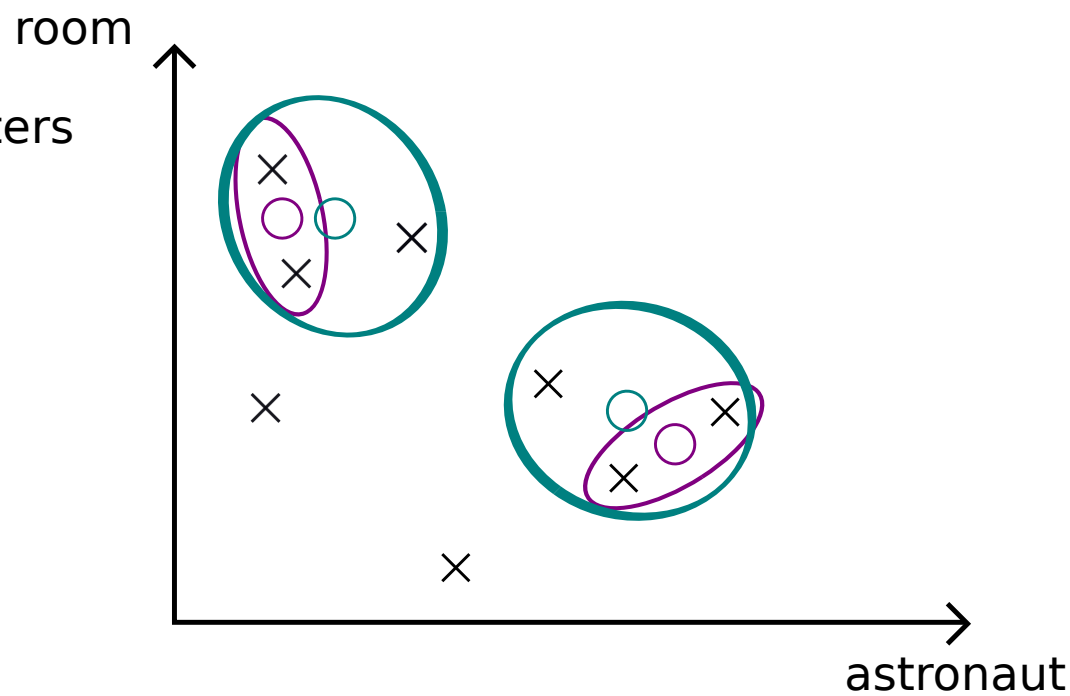


Pre-Clustering:

Group Average Agglomerative Clustering (GAAC)

5/7

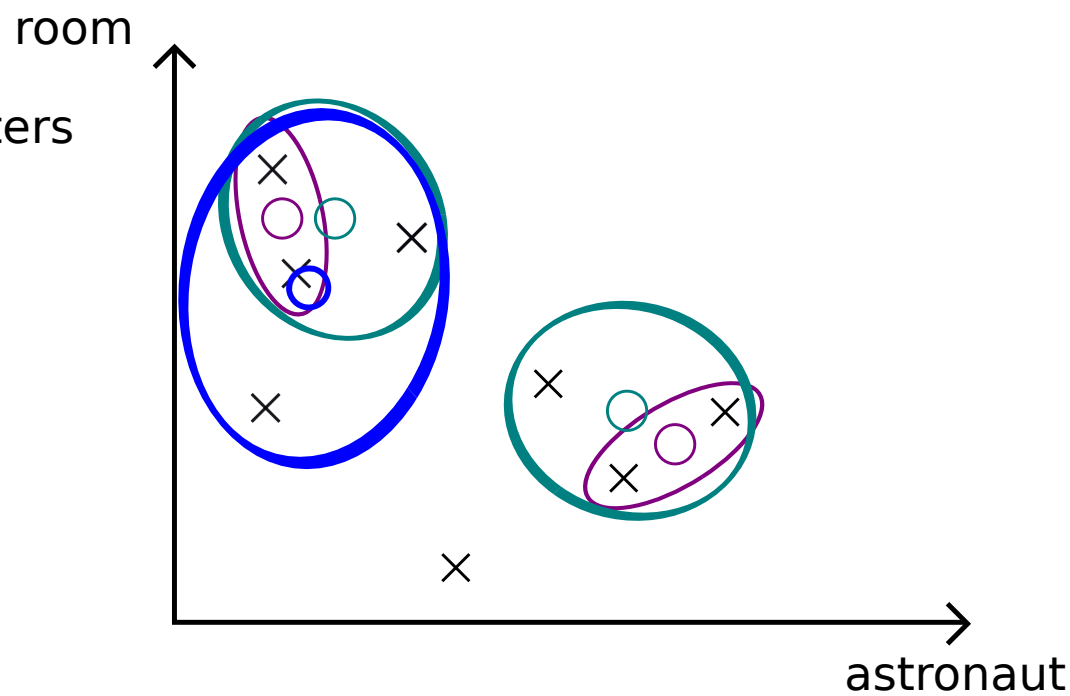
- unlabeled context vectors
- initially all datapoints are clusters
- repeatedly merge closest cluster pair (closest average distance = closest averages = closest centroids)



Pre-Clustering:

Group Average Agglomerative Clustering (GAAC)
6/7

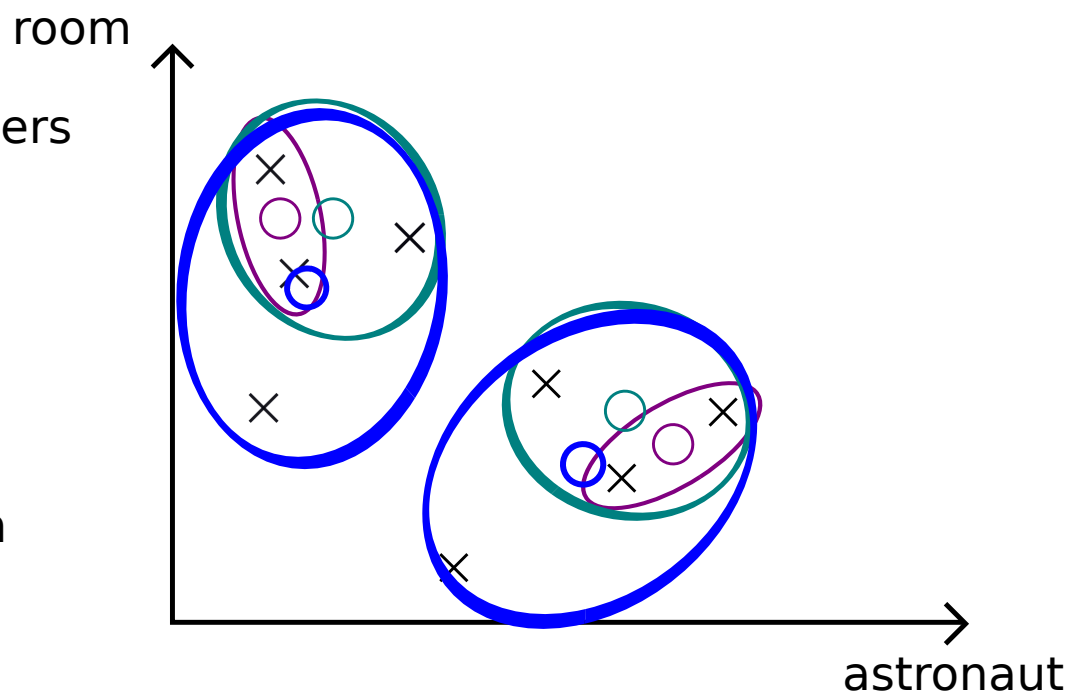
- unlabeled context vectors
- initially all datapoints are clusters
- repeatedly merge closest cluster pair (closest average distance = closest averages = closest centroids)



Pre-Clustering:

Group Average Agglomerative Clustering (GAAC) 7/7

- unlabeled context vectors
- initially all datapoints are clusters
- repeatedly merge closest cluster pair (closest average distance = closest averages = closest centroids)
- usually better cluster results
- problem: $O(N^2)$
- GAAC preclustering on random sample. Use result as intitial k-Means cluster centroids



Evaluation

- Dataset

  17 months of the New York Times News Service

  June 1989 through October 1990. (about 435 megabytes and 60.5 million words)

  Two months reserved as test set

  November 1990 and May 1989 (46 megabytes, 5.4 million words)

- 10 natural ambiguous words with hand-labeled senses in the test set

- trick: 10 artificial ambiguous *words*

  An ambiguous word is a new word, that replaces two natural (combined) words (e.g. *wide range / consulting firm*. The task is to find the original replaced word.

Automatic Word Sense Discrimination

Results

accuracy on the test set

Word	Train	Test	% Rare Senses	% most freq	local, k=2
wide range / consulting firm	1,422	149	0	62	56
heart disease / reserve board	1,197	115	0	54	87
space	9,136	200	0	56	61
tank	3,909	200	1.5	90	95
average			2.1	61.2	77.8

space: area, volume / outer space

tank: combat vehicle / receptacle for liquids

Number of clusters

- k=2: *outer space* and *limited extent in one, two, or three dimensions*
- k=10: *office space*, *exhibition space*, *disk space*, ...
- evaluation problem: unclear/subjective which sense labels to assign - overlapping senses
- indirect evaluation
  - homogenity of the clusters w.r.t. the k=2 problem
  - improvement of IR task by using sense dimensions instead of word dimensions

Results

accuracy on the test set

Word	Train	Test	% Rare Senses	% most freq	local, k=2	local, k=10
wide range / consulting firm	1,422	149	0	62	56	64
heart disease / reserve board	1,197	115	0	54	87	85
space	9,136	200	0	56	61	71
tank	3,909	200	1.5	90	95	91
average			2.1	61.2	77.8	81.8

space: area, volume / outer space

tank: combat vehicle / receptacle for liquids

Singular Value Decomposition 1/2

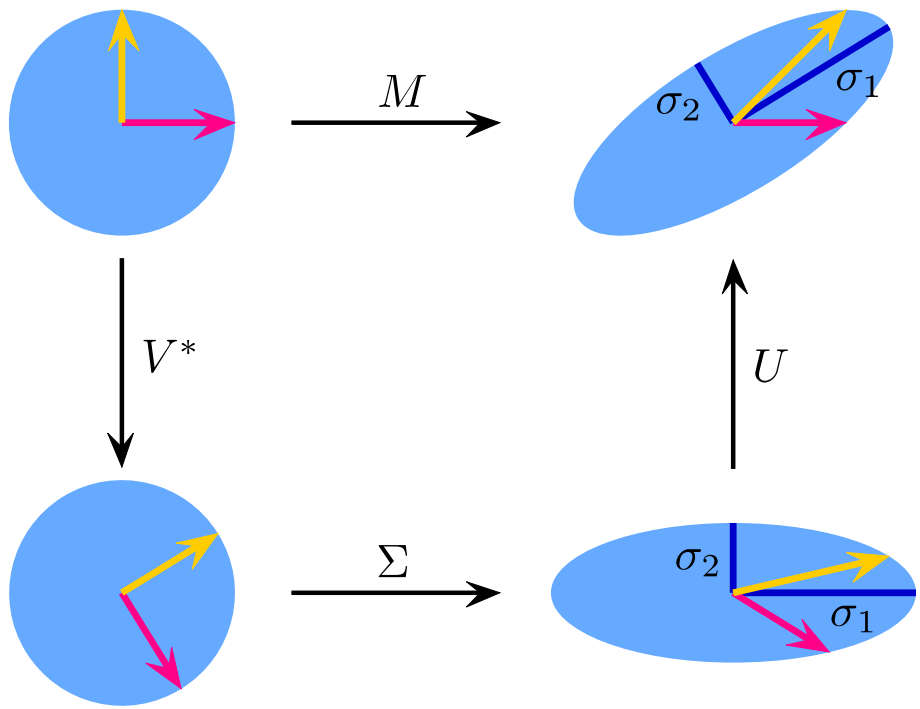
- transforming a first-order co-occurrence vector $\vec{x}$ to a second-order co-occurrence vector $\vec{v}$ can be represented by:

$$\vec{v} = M_{d \times f} \ \vec{x}$$

  where M is a d-by-f matrix with d being the number of dimensions in the vector-space and f being the number of word vectors (or *features*). The columns in M are exactly the word vectors.

- $M_{d \times f}$ can be decomposed into $M_{d \times f} = U_{d \times p} \ \Sigma_{p \times p} \ V^*_{p \times f}$
- $U_{d \times p}$ and $V^*_{p \times f}$ are orthonormal matrices. These do not change the value of dot product of any two vectors transformed by the matrix.
- $\Sigma_{p \times p}$ is a diagonal matrix, that only scales coordinates.

Singular Value Decomposition 2/2



$$M = U \cdot \Sigma \cdot V^*$$

source: Wikipedia (Creative Commons)

Dimensionality reduction by keeping only the most scaled N dimensions in Σ

Instead of $\vec{v} = M_{d \times f}\ \vec{x} = U_{d \times p}\ \Sigma_{p \times p}\ V^*_{p \times f}\ \vec{x}$ we write $\vec{v}' = \Sigma'_{p' \times p'}\ V^*_{p' \times f}\ \vec{x}$

Results

accuracy on the test set

Word	% most freq	local, k=2	local, k=10	local, SVD, k=2	local, SVD, k=10
wide range / consulting firm	62	56	64	65	66
heart disease / reserve board	54	87	85	99	99
space	56	61	71	60	76
tank	90	95	91	87	92
average	61.2	77.8	81.8	82.9	88.3

space: area, volume / outer space

tank: combat vehicle / receptacle for liquids

Results

accuracy on the test set

Global feature selection:

- 2,000 most frequent words in the complete training data as dimensions
- 20,000 most frequent words in training data as features (word-vectors)

Word	% most freq	local, k=2	local, k=10	local, SVD, k=2	local, SVD, k=10	global, SVD, k=2	global, SVD, k=10
wide range / consulting firm	62	56	64	65	66	87	87
heart disease / reserve board	54	87	85	99	99	100	100
space	56	61	71	60	76	56	75
tank	90	95	91	87	92	85	92
average	61.2	77.8	81.8	82.9	88.3	89.7	90.6

Conclusion

- a completely unsupervised method
- level of detail is configurable via the number of clusters
- all average results above baseline
- best variant (SVD on global feature-selection, k=10) still about 5-10 % lower accuracy than Yarowskys WSD
- different task
  - easier? - we do not need to decide on a (given) label
  - harder? - we not only decide between given groups but also induce them

Thanks for your attention!

questions

discussion