Clustering Semantically Similar words

PART I: Semantically Similar words

PART II: Clustering

Motivation (from QA)

- Question Classification

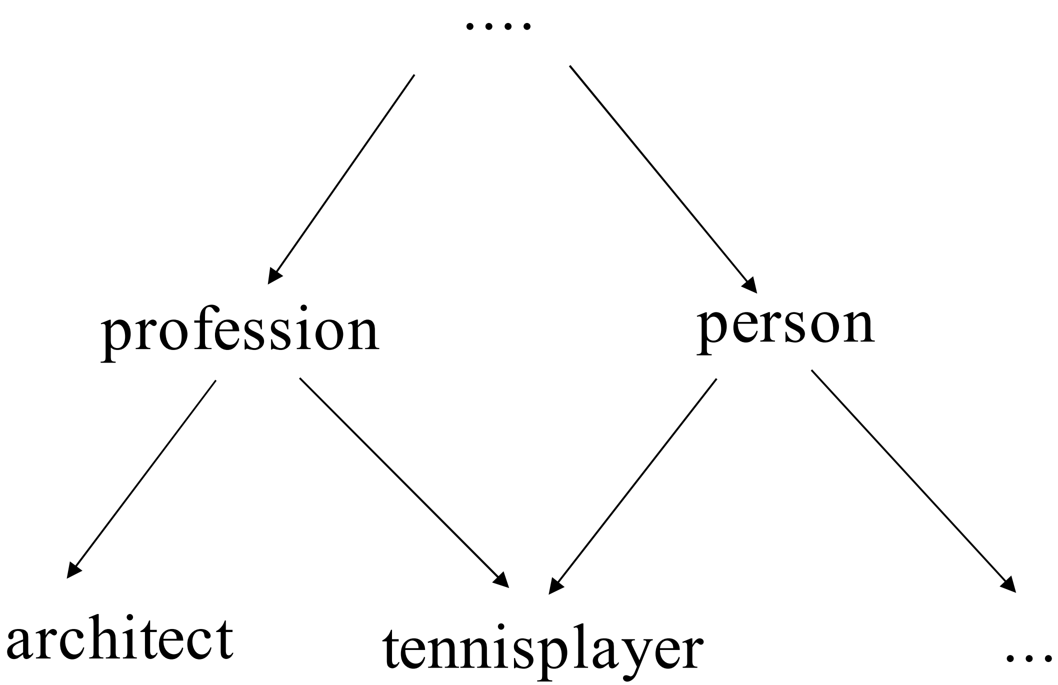
`Welke tennisser ...?` > person ques.

- Answering 'which' questions

`Welk beroep heeft Renzo Piano?`

`De` **`Italiaan`** `Renzo Piano is` **`architect.`**

What do we need?



Motivation(from parsing)

- Disambiguation
  - Coordination:

    (Chirac uit Frankrijk) en Blair.

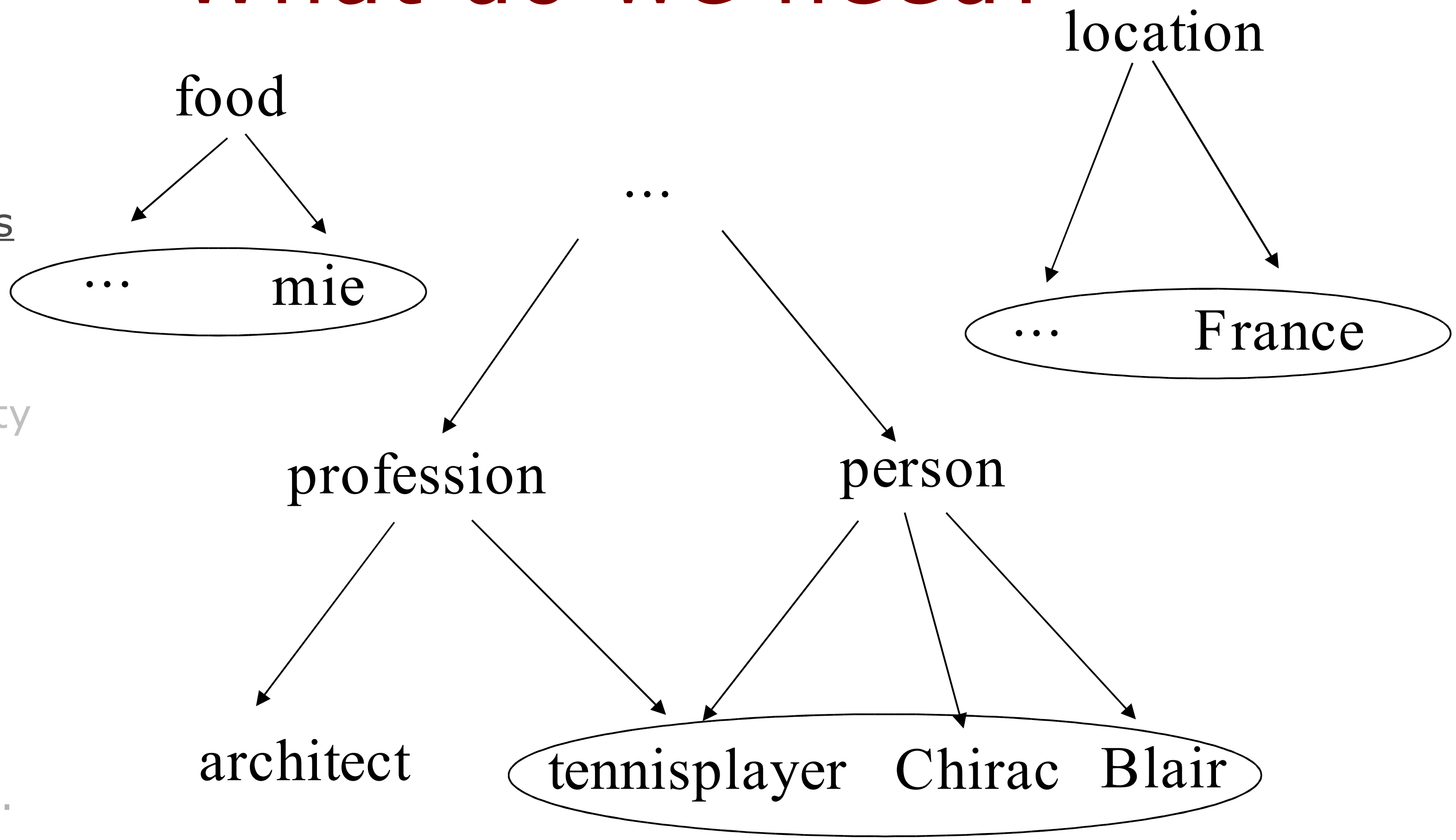
    Chirac uit (Frankrijk en Blair).

  - PP-attachment

    Hij at (mie) met stokjes.

    Hij at (mie met stokjes).

What do we need?



Dutch EuroWordNet

- There is a resource for Dutch
- However, its coverage is not sufficient for our task.
- From 72 function questions, 21 missclassified, because of missing functions in EWN

Distributional Similarity

- Semantically similar words share similar contexts
- Context = Syntactic context

	zie	verf	verzorg	laat_uit
bus	50	5	1	0
hond	56	1	5	8

Data

- 78 million words of parsed Dutch newspaper text

Subject-Verb	kat eet
Verb-Object	voer kat
Adjective-Noun	langharige kat
Coordination	Bassie en Adriaan
Apposition	de clown Bassie
PC	begin_met werk

Extracted from data

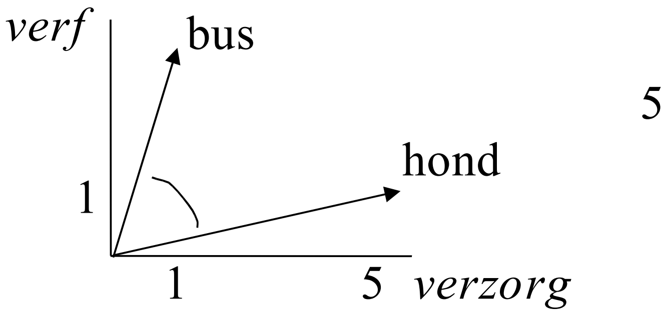
Gram rel.	# tuples
Subj	5.639.140
Adj	3.262.403
Obj	2.642.356
Coord	965.296
PC	770.631
Appo	602.970

	ga_subj	geel_adj	neem_obj	Lassie_app
bus	4	9	8	0
hond	4	1	6	8

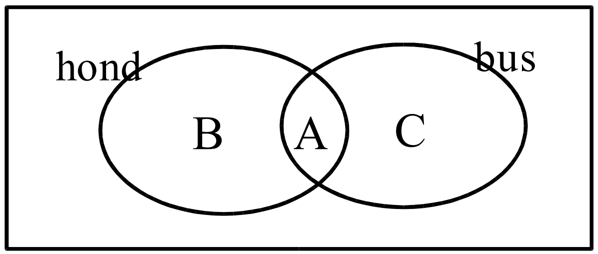
Cutoff:
row > 10

Similarity Measures

- Cosine



- Dice



$$\frac{2A}{2A+B+C}$$

Weights

- Mutual Information

	hebben	verdoen	geven	doordrijven
zin	500	0	400	18
tijd	560	10	600	0

Example Output

danser:

muzikant,
musicus,
choreograaf,
danseres,
artiest,
renner,
zanger,
personage,
orkest, cabaretier, sporter, bezoeker, acteur,
vrijwilliger, theatermaker, technicus, clown,
toneelspeler, politiemens, actrice,

Demo: http://www.let.rug.nl/vdplas_bin/verwant.py

Example Output

Michael Jackson:

Prince,
Tina Turner,
Elton John,
Madonna,
Peter Gabriel,
Bruce Springsteen,
Genesis,
Rolling Stones,
Dire Straits,
David Bowie, U2, The Rolling Stones, Paul Simon, Simple Minds, Beatles, Elvis Presley, Sting, Bob Dylan, Bon Jovi, Neil Young,

Example Output

onvrede:

ongenoegen,
ontevredenheid,
frustratie,
bezorgdheid,
ergernis,
ongerustheid,
verontwaardiging,
boosheid,
irritatie,
weerzin, wantrouwen, verwarring, onrust,
onzekerheid, teleurstelling, afkeer, vrees,
onbehagen, scepsis, twijfel,

Example output

blad:

tijdschrift,
krant,
weekblad,
dagblad,
maandblad,
medium,
pers,
magazine,
blaadje,
bloem,
televisie, stuk, tak, steen, materiaal, plant,
stukje, artikel, boekje, schilderij,

Evaluation Framework

- 1000 words from EWN (random, freq> 10)

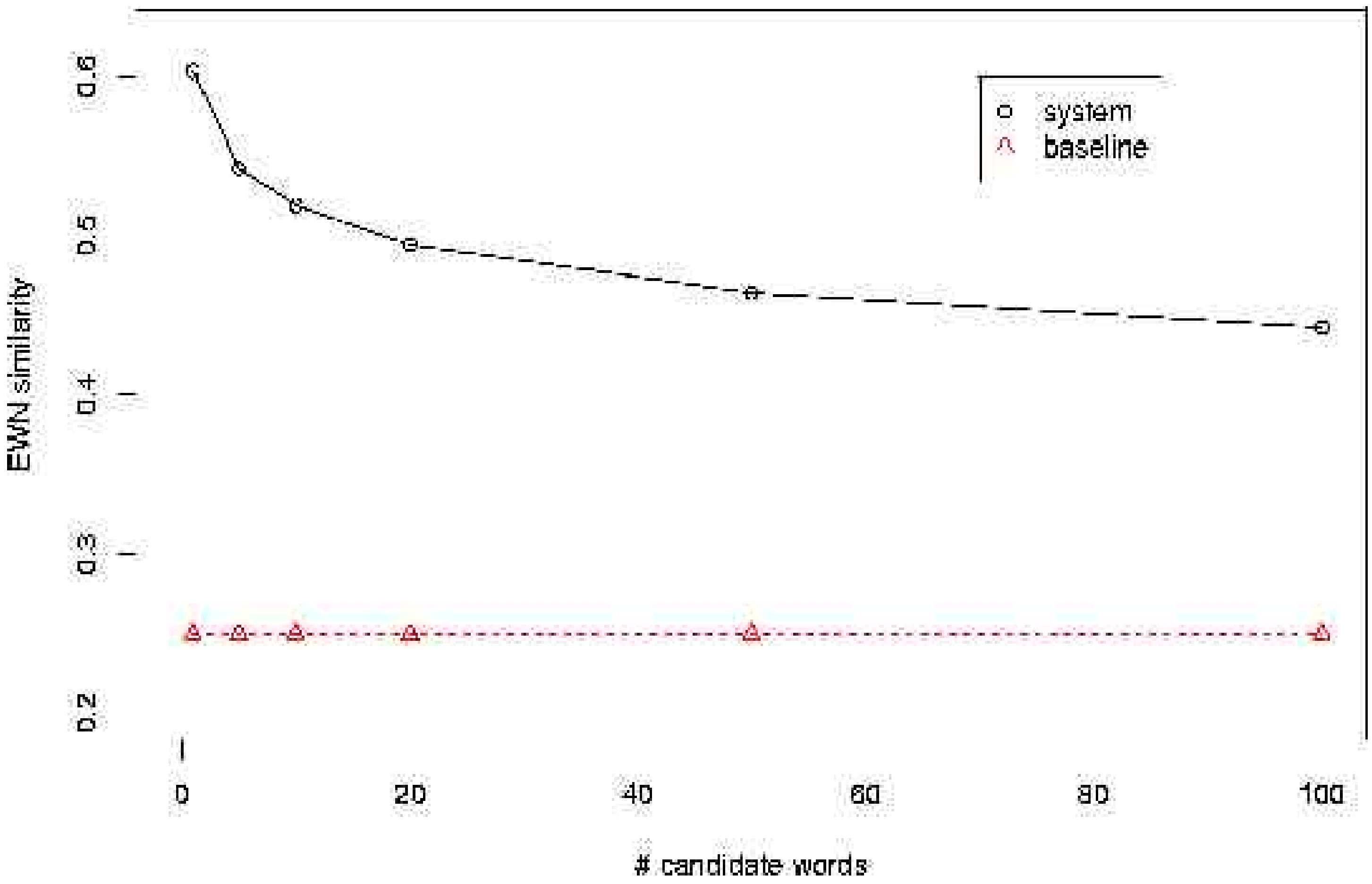
→ For each word

collect 100 most similar words according to system

→ For each pair of words

measure similarity in EWN using Wu&Palmer measure(1994)

Performance



PART I :Similar words

Motivation

Distributional similarity

Data

Similarity Measures

Output

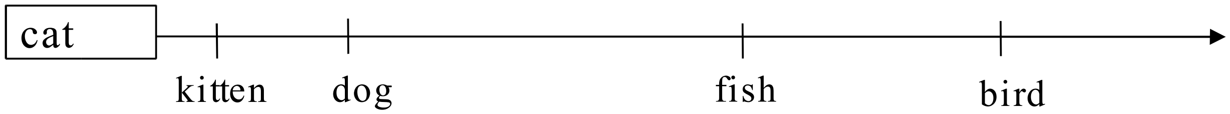
Evaluation + perform.

Conclusion

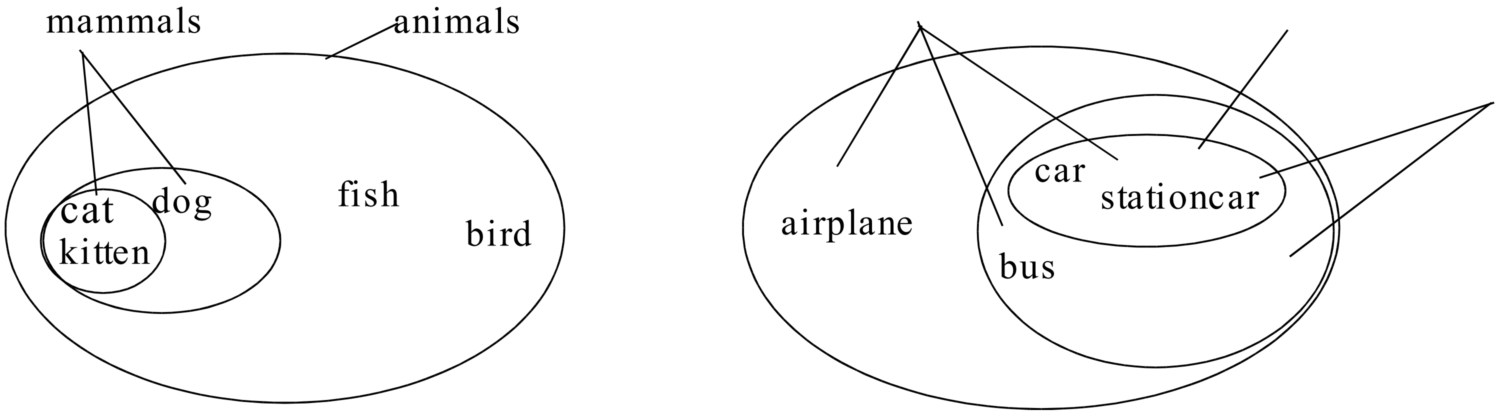
PART II Clustering

Conclusions

- We now gave as output a ranked list of semantically similar words for each word



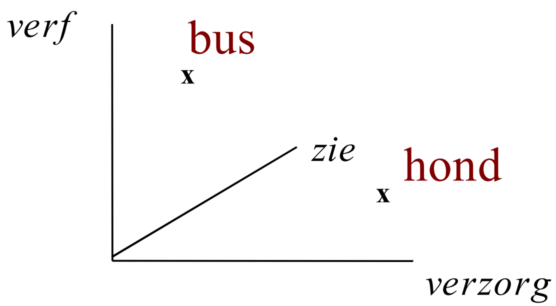
- We want to look at the groups that can be formed (semantic classes)



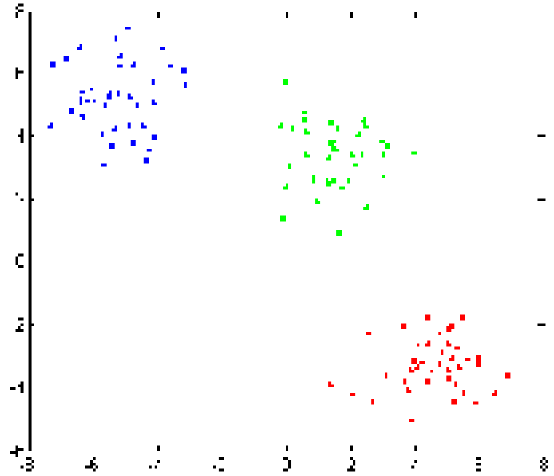
Conclusions(cntd.)

- We have as data:

	zie	verf	verzorg	laat_uit	
bus	50	5	1	0	etc
hond	56	1	5	8	etc
	etc	etc	etc	etc	



Clustering



- Clustering algorithms partition a set of objects into groups or clusters.
- The criterion for clustering is similarity.
- It's like putting your dirty clothes in piles of similar colours and similar washing instructions.

Motivation

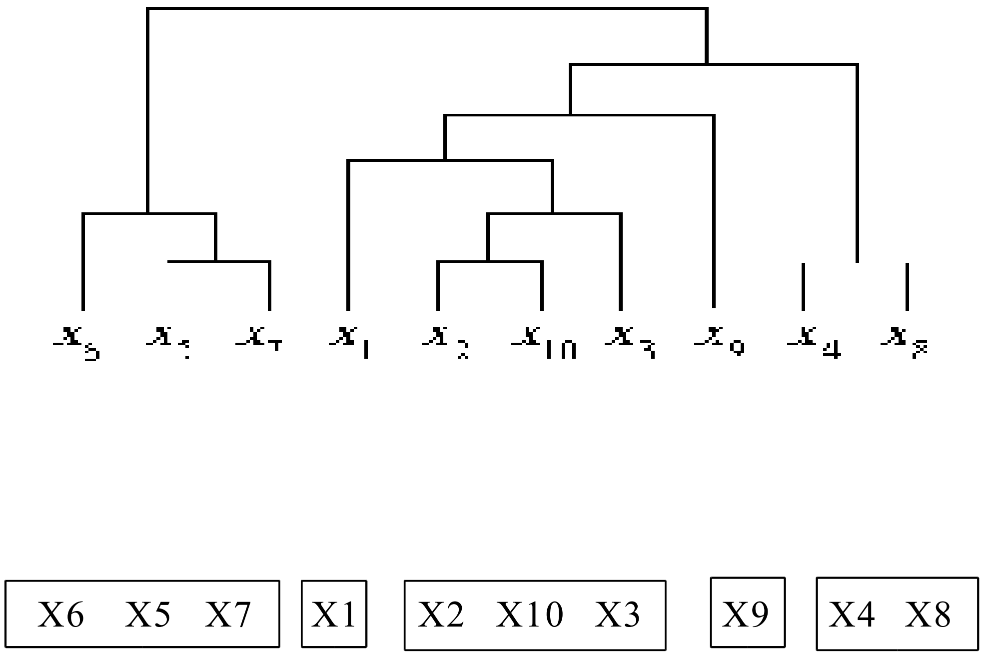
- Exploratory data analysis

Clustering people with language impairments according to the similarity in mistakes they make and see what groups emerge after clustering

- Generalization

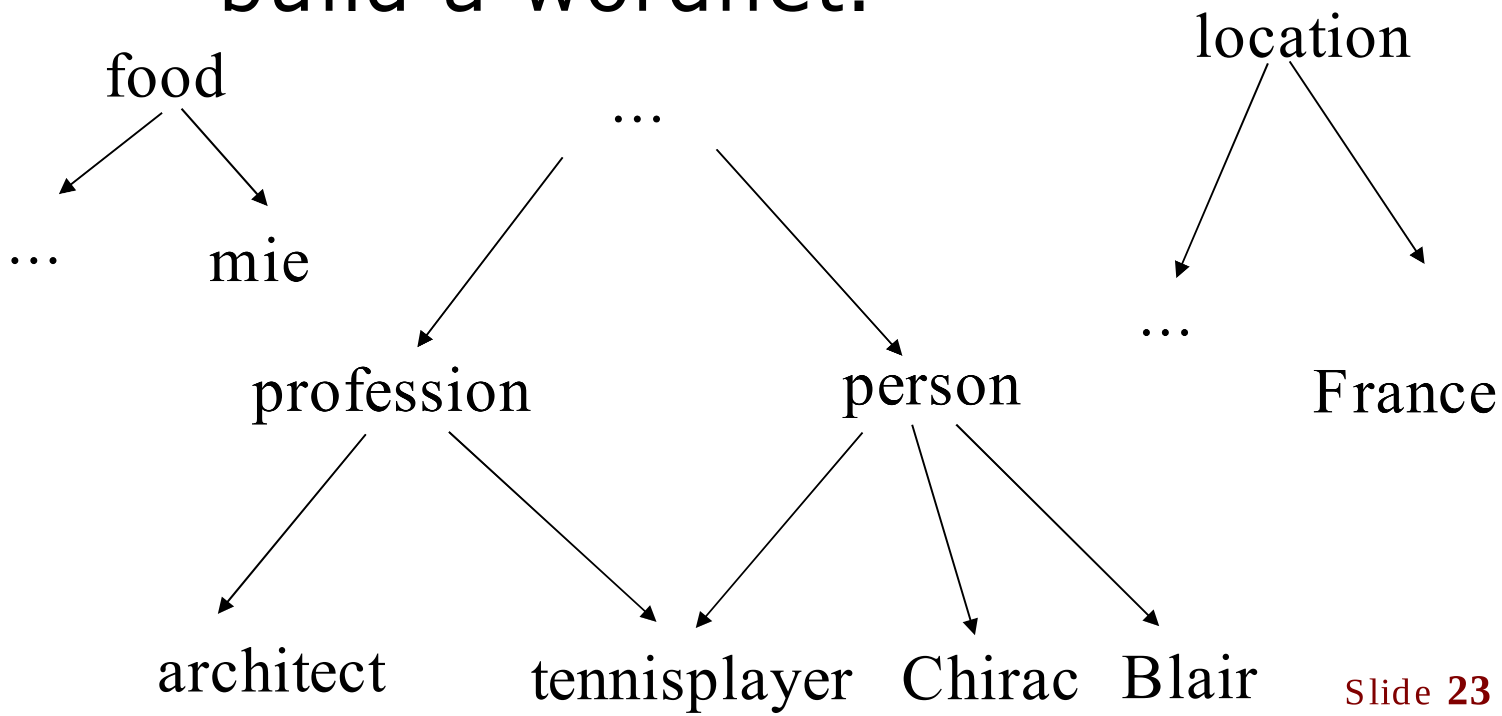
All kinds of animals will be in one cluster despite their differences

Hierarchical versus flat



Hierarchical versus flat (cntd)

- It seems that hierarchical clustering is what we need to build a wordnet.



Will hierarchical clustering give us the hierarchy we want?

danser:

muzikant,

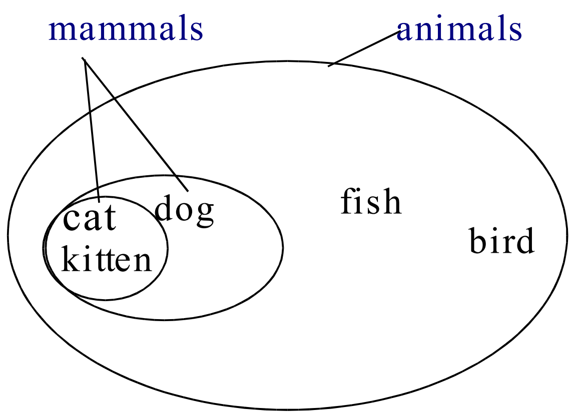
musicus,

choreograaf,

danseres,

artiest,

renner, zanger, personage, orkest, cabaretier, sporter, bezoeker, acteur, vrijwilliger, theatermaker, technicus, clown, toneelspeler, politiemens, actrice,



Labels??

PART I :Similar words

PART II Clustering

Intro

hierarchical vs flat

soft vs hard

clash + solutions

flat clustering(hard)

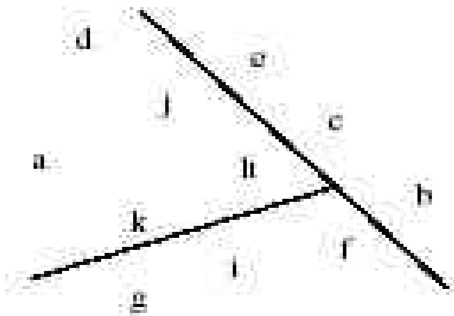
soft clustering

more about senses

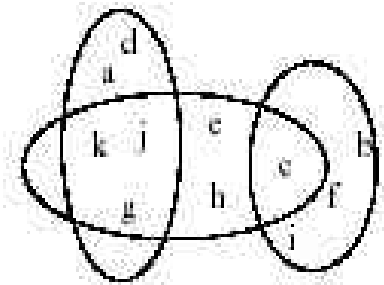
Soft versus hard

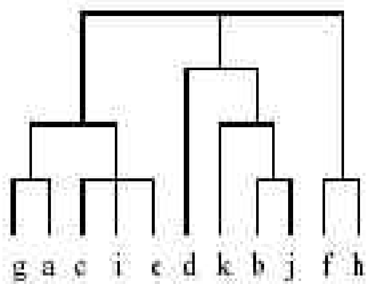
- hard: Each object is assigned to only one cluster.
- Soft clustering allows degrees of memberschip and membership in multiple clusters as a degree in certainty.
- Disjunctive models: true multiple assignment (not just uncertainty)

Examples



	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1





Soft versus hard (cntd)

- Language is not unambiguous and hard clustering does not seem a good way to deal with linguistic data.
- In my case polysemous words ('blad'). They belong to more than one cluster.

Clash

- In hierarchical clustering assignment is usually hard.

- In flat clustering: hard or soft

Possible solutions

- 1. first soft/disjunctive clustering
- 2. Give senses a unique ID, split-up features
- 3a. do hard/hierarchical clustering with senses included
- 3b. get hierarchical information from other source

Other sources for hierarchical information

- Patterns in free text such as
  Xs, Ys and other Zs

  Zs , such as Xs and Ys

  But can be subjective:
  Balkenende, Rutte and other disasters.
- Dictonaries, encyclopedia
- Dutch EWN

Flat clustering (term.)

- Cluster centre = centre of the M points in a cluster c = centroid
- Each component of the centroid vector is simply the average of the values for that component

	zie	verf	verzorg	laat_uit	
bus	50	5	1	0	…
hond	56	1	5	8	…
Centroid	53	3	3	4	…

Flat clustering (hard)

- Set of initial cluster centres
- Go through several iterations:

- of assigning each object to closest cluster centre.
- and recomputing the cluster centre

- Repeat until stable

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1

Soft clustering

PART I :Similar words

PART II Clustering

Intro

hierarchical vs flat

soft vs hard

clash + solutions

flat clustering (hard)

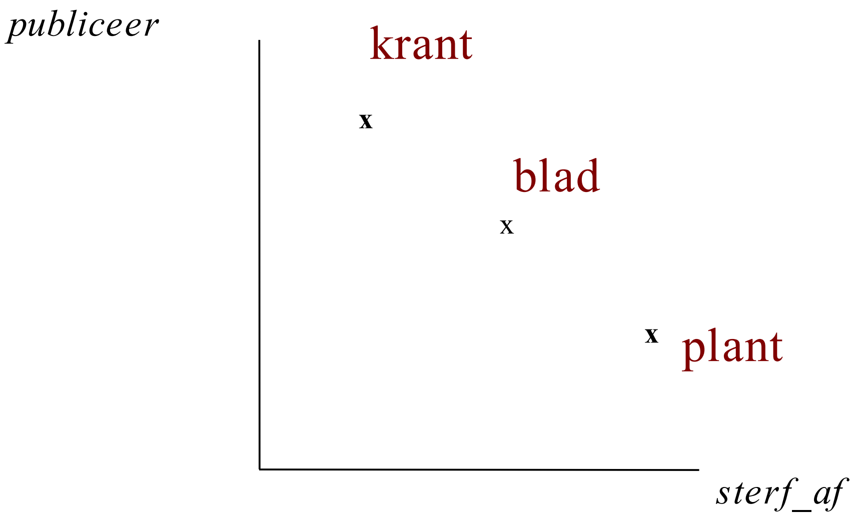
soft clustering

more about senses

- The idea is that the observed data are generated by several causes = clusters
- Gaussian mixture model: For each element the clusters from flat clustering are still the dominant clusters. But each word also has some non-zero membership in other cluster.

More about senses

- Example *blad*:



PART I :Similar words

PART II Clustering

Intro

hierarchical vs flat

soft vs hard

clash + solutions

flat clustering (hard)

soft clustering

more about senses

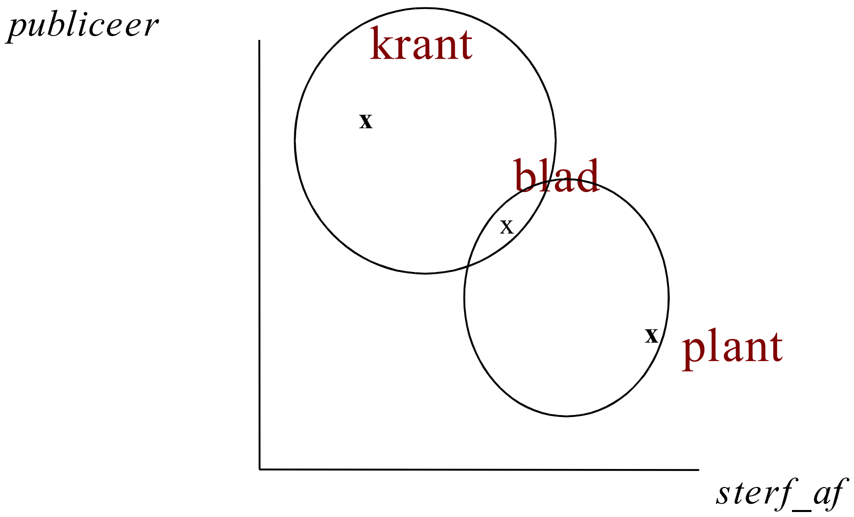
More about senses

- Soft clustering:

	1	2	...
Krant	0.9	0	...
Blad	0.6	0.5	...
Plant	0	0.9	...

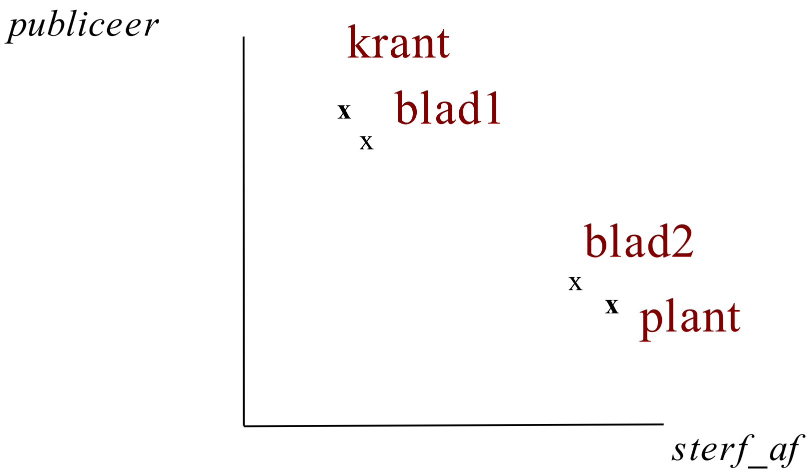
More about senses

- Disjunctive clustering



More about senses

- What I really want:



Use bidirectionality

- Recap: Distributional similarity=>Semantically similar words share similar contexts
- Our goal was clusters of nouns and verbs were our features
- Nouns can become features and verbs can become elements to be clustered.

Using this bi-directionality to get senses

- Would it be possible to first cluster the features(verbs)
- split nouns up according to number of clusters found in features (verbs) >senses
- and give each sense its accompanying subset of features.

PART I :Similar words

PART II Clustering

Intro

hierarchical vs flat

soft vs hard

clash + solutions

flat clustering (hard)

soft clustering

more about senses

PART I :Similar words

PART II Clustering

Intro

hierarchical vs flat

soft vs hard

clash + solutions

flat clustering (hard)

soft clustering

more about senses

Example: blad

- publiceer, geef_uit, distribueer, hark, kauw, sterf_af
- Decide based on context of these verbs that they can be split in two clusters
- Cluster 1:publiceer, geef_uit, distribueer
- Cluster 2: hark, kauw, sterf_af
- => Cluster1 becomes blad#sense1
- => Cluster2 becomes blad#sense2

PART I :Similar words

PART II Clustering

Intro

hierarchical vs flat

soft vs hard

clash + solutions

flat clustering (hard)

soft clustering

more about senses

Thank you!

Hierarchical vs flat

PART I :Similar words

PART II Clustering

Intro

hierarchical vs flat

soft vs hard

clash + solutions

flat clustering(hard)

soft clustering

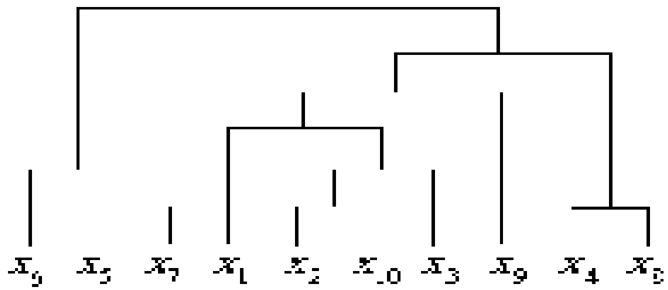
more about senses

Hierarchical

- More info
- Less efficient
- No single best algorithm, depends on task

Flat

- Less info
- More efficient
- K-means, but assumes Euclidean space , not good for nominal data (nor for probabilities?) > EM algorithm based on probabilistic models



Bottom-up vs top-down

- You start with every object in one cluster en start merging them

- You start with one big cluster and look for weakest link. The least coherent cluster is split.

Bottom-up vs top-down

- Splitting clusters is a clustering task in itself. > choose for bottom-up (iteration)
- However, our objects have many zero's in prob. distr. and bottom-up cannot handle that.