- Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30(1-7), 107–117.
- Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research (JAIR), 22, 457–479.
- Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into texts. In Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP-04), pages 404–411, Barcelona, Spain.
- Mihalcea, R. and Radev, D. (2011). Graph-based Natural Language Processing and Information Retrieval. Cambridge University Press

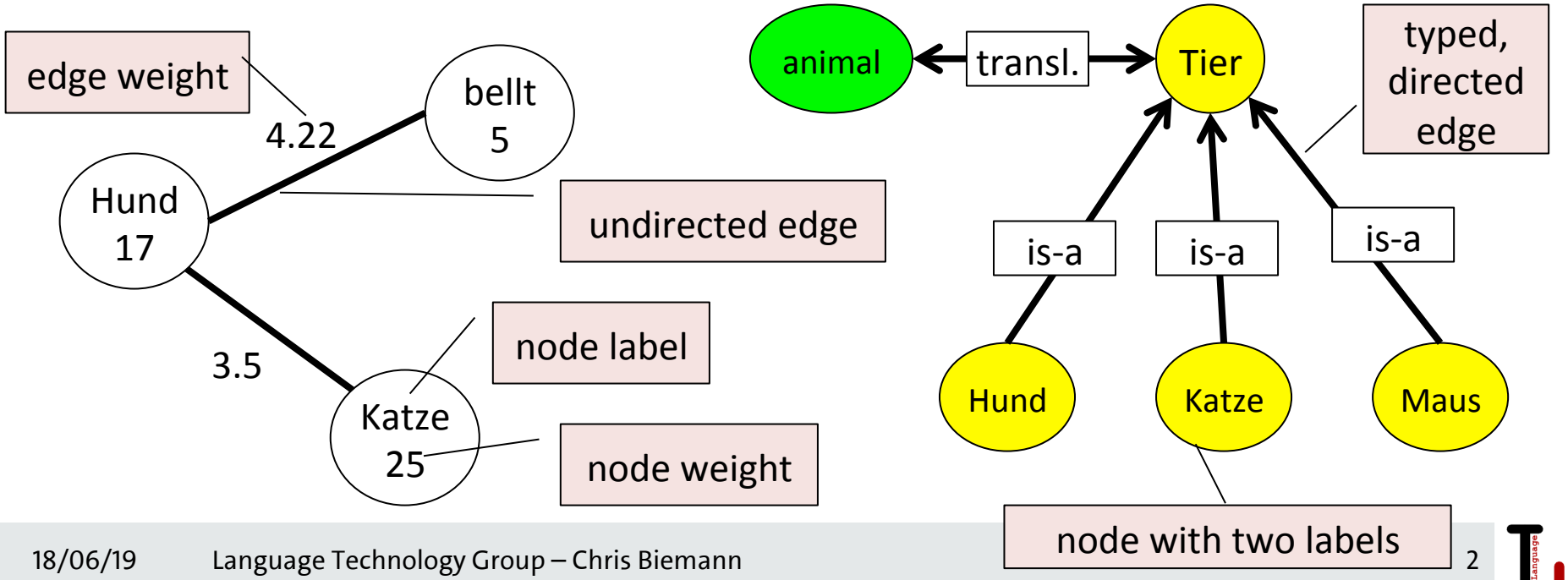
graphs, networks, PageRank, graph clustering

GRAPH-BASED METHODS FOR NLP

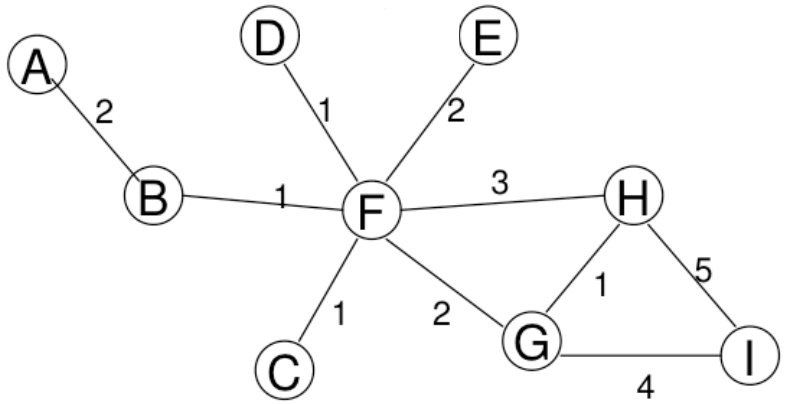
GRAPHS IN NLP

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Some (language) data is naturally represented as a graph
- Graph G(V,E) with V = set of nodes, E = set of edges
  - nodes can be (multi-)labeled, weighted
  - edges can be weighted, typed, directed or undirected
- Graphs encode relationships between entities, which are represented as nodes



GRAPH REPRESENTATIONS



Geometrical representation

a_{ij}	A	B	C	D	E	F	G	H	I
A	0	2	0	0	0	0	0	0	0
B	2	0	0	0	0	1	0	0	0
C	0	0	0	0	0	1	0	0	0
D	0	0	0	0	0	1	0	0	0
E	0	0	0	0	0	2	0	0	0
F	0	1	1	1	2	0	2	3	0
G	0	0	0	0	0	2	0	1	4
H	0	0	0	0	0	3	1	0	5
I	0	0	0	0	0	0	4	5	0

Adjacency matrix

Set representation

G(V,E)

V={A,B,C,D,E,F,G,H,I}

E={AB,BF,CF,DF,EF,FG,FH,GH,GI,HI} }

ew(BF)=ew(CF)=ew(DF)=ew(GH)=1;

ew(AB)=ew(EF)=ew(FG)=2; ew(FH)=3;

ew(GI)=4; ew(HI)=5;

MOTIVATION FOR GRAPH REPRESENTATION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Graphs are an intuitive and natural way to encode language units as nodes and their similarities as edges
  - feature-based representation can be transformed into a graph via a similarity measure
  - graphs cannot be transformed back into the feature representation.
  - Think e.g. points in space
- There exist methods that directly operate on graphs
  - graph clustering, e.g. MinCut
  - graph-based ranking, e.g. PageRank
  - path-based algorithms, e.g. Dijkstra
- Representation as adjacency matrix A: A(i,j)=weight of edge (i,j)

PAGE RANK FOR WEB PAGES

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- First-generation Google global ranking algorithm (1998)
- Measure the (query-independent) importance of Web page based on the link structure alone.
- Assign each node a numerical score between 0 and 1: PageRank.
- Rank Web pages based on PageRank values.

Idea:

- Every page has a number of in-links (back links) and out-links (forward links)
- pages with more in-links are more important
- in-links from important pages are more important
- Back in 1998, there were no link farms.

Charles H. Hubbell: An input-output approach to clique identification. In: Sociometry 28, pp. 377-399, 1965

DEFINITION OF PAGERANK

u: a web page. R_u its page rank

F_u: set of pages u points to

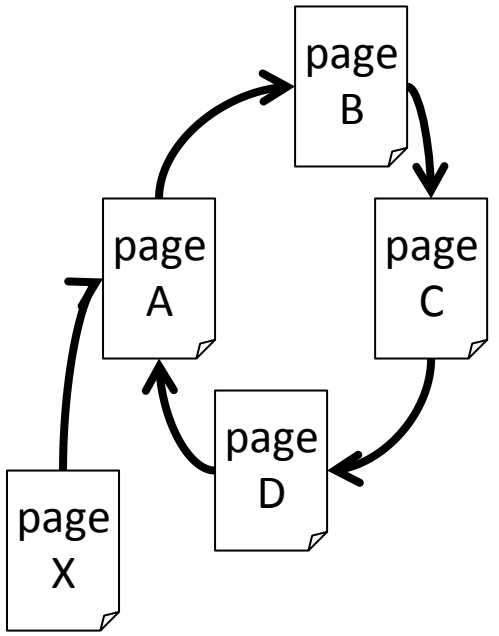
B_u: set of pages that point to u

$|F_u|$: the number of links from u

N: total number of pages

d: damping factor, default d=0.85

$$R(u) = \sum_{v \in B_u} \frac{R(v)}{|F_v|} \cdot d + \frac{(1-d)}{N}$$

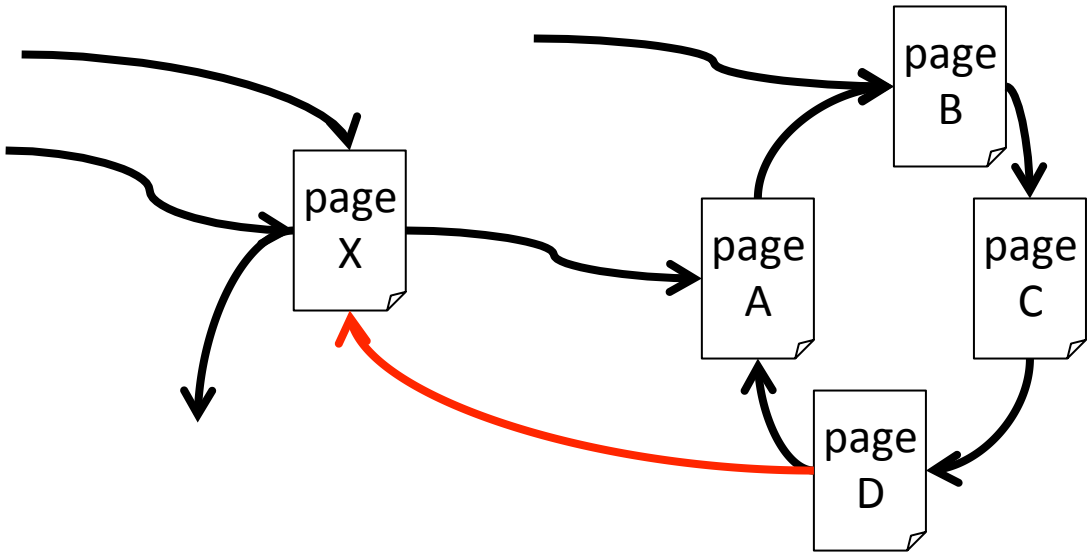


- The equation is recursive, but it may be computed by starting with any set of ranks and iterating the computation until it converges.
- Rank sink problem: ring of pages that accumulated rank, but never distributes rank
- Need damping: uniform rank distribution for all pages

RANDOM SURFER MODEL

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- When normalizing PageRank over all pages, *R(u)* can be thought of as the probability that a random surfer looks at page *u*.
- Damping corresponds to "teleportation": With some probability *d*, the random surfer is teleported to some other page



MATRIX NOTATION

$$R(u) = \sum_{v \in B_u} \frac{R(v)}{|F_v|} \cdot d + \frac{(1-d)}{N}$$

- PageRank can be written as
  $\mathbf{R}=(1\text{-}d)\mathbf{D} + [d\mathbf{B}]^T\mathbf{R}$ where
  - **D** is a vector of dimension N where all elements are equal to 1/N
  - **B** is a a stochastic version of the adjacency matrix $B(i,j) = \frac{A(i,j)}{\sum_k A(i,k)}$
- Stationary distribution: If a stochastic matrix **X** is irreducible and aperiodic, there exists a unique stationary distribution **r**
  - irreducible: path between all nodes
  - aperiodic: no rank sinks

  $\lim_{n \to \infty} \mathbf{X}^n = \mathbf{1}^T\mathbf{r}$
- Teleportation with damping factor makes the effective transition matrix irreducible and aperiodic!

COMPUTATION OF PAGERANK

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Numeric: Simulate a lot of random surfers: The Power method of Eigenvector computation
  - initialize all pages with the same rank
  - repeat until convergence:
    - for all pages u: compute $R_{t+1}(u)$ on the basis of $R_t(v)$
    - t:=t+1

input : matrix size N , error tolerance ϵ
output: eigenvector p

$p_0 = 1/N\ 1$
t=0;
repeat until $\delta < \epsilon$:
  t=t+1;
  $p_t = M^T p_{t-1}$;
  $\delta = ||p_t - p_{t-1}||$;
return p_t ;

LexRank: Application to Multi-Document Summarization

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Multi-document summarization:
  1. identify important topics of the documents to be summarized
  2. identify sentences belonging to a certain topic
  3. from sentences belonging to the same topic, select the ones that best describe the topic
  4. concatenate sentences from different topics and make sure they fit together

- Sub-problem 3.:
  - input: Sentences that talk about more or less the same thing
  - output: scores for those sentences that reflect how well a single sentence represents that topic
- Idea: use measures on sentence similarity graph

FROM SENTENCES TO TF*IDF VECTORS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

	TF: count $w_1..w_n$				TF*IDF			
Sentence	**w1**	**w2**	**w3**	**wn**	**w1**	**w2**	**w3**	**wn**
This is a sentence that talks about some topic.	3	0	2	0	0	0	.27	0
And here is another sentence that talks abot something slightly different.	5	0	3	1	0	0	.24	.21
And here is yet another one of these notorious sentences	7	4	0	0	0	.62	0	0
DF	**3**	**1**	**2**	**1**				

feature vector of the second sentence

This is the same as the vector space model for Information Retrieval

$$IDF(w) = \log\left(\frac{\text{total number of sentences}}{DF(w)}\right)$$

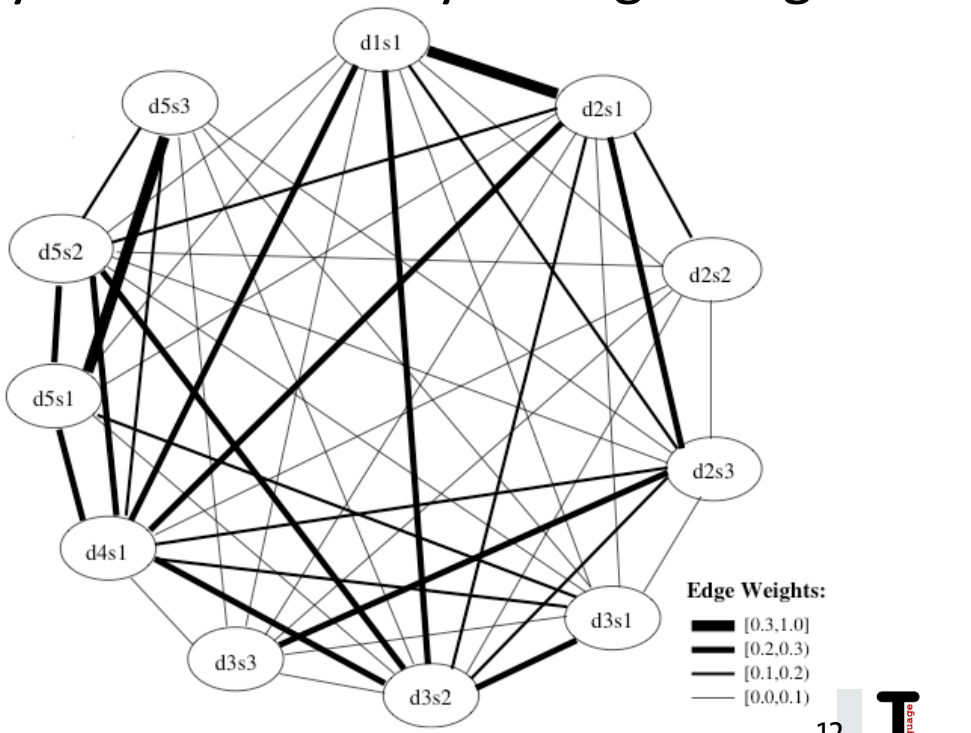
FROM TF*IDF VECTORS TO SENTENCE SIMILARITY GRAPH

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Sentence similarity graph:
  - nodes: sentences
  - edges: cosine similarity between sentence feature vectors

$$\frac{\sum_{w \in x,y} \mathrm{tf}_{w,x} \mathrm{tf}_{w,y} (\mathrm{idf}_w)^2}{\sqrt{\sum_{x_i \in x} (\mathrm{tf}_{x_i,x} \mathrm{idf}_{x_i})^2} \times \sqrt{\sum_{y_i \in y} (\mathrm{tf}_{y_i,y} \mathrm{idf}_{y_i})^2}}$$

- Can apply threshold on similarity or use similarity as edge weight
- Could use any other method that produces sensible sentence vectors



CENTROID, DEGREE AND CENTRALITY

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Centroid
  - idea: select average sentence. Compute average point of sentence vectors (centroid)
  - select sentence that is most similar to the centroid for summarization
- Degree Centrality
  - idea: sentences that cover most of the content have a high node degree (number of edges): since word overlap is responsible for edges, degree measures word overlap with the overall set of sentences
  - for summarization, choose the sentence with the highest degree
- LexRank Centrality
  - idea: it does not suffice to be similar to many sentences: similarity to important sentences counts more.
  - normalize the adjacency sentence similarity to make it a stochastic matrix
  - run PageRank to obtain scores that are used for ranking the sentences
  - for summarization, choose sentence with highest score

EVALUATION OF GRAPH-BASED MULTI-DOCUMENT SUMMARIZATION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

	2004 Task4a		
	min	max	average
Centroid	0.3768	0.3901	0.3826
Degree (t=0.1)	0.3863	0.4027	0.3928
LexRank (t=0.1)	0.3931	0.4038	0.3974
Cont. LexRank	0.3924	0.4002	0.3963

baselines: random: 0.3593
lead-based: 0.3788

- Scores: ROUGE metric: similar to BLEU, between manual summaries and system summaries
- random baseline: select any sentence from set by chance
- lead-based: select based on position of sentence within document

➔ LexRank is a simple method for getting high scores. It uses the whole structure of the graph, as opposed to Centroid or Degree.

This technique also works well for single-document summarization.

CO-OCCURRENCE GRAPHS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- If two words X and Y occur together in some contextual unit of information (as neighbors, in a word window of 5, in a clause, in a sentence, in a paragraph), they are said to co-occur.
- When regarding words as vertices and edge weights as the number of times two words co-occur, the word co-occurrence graph of a corpus is given by the entirety of all word co-occurrences
- To find out whether the co-occurrence of two specific words A and B is merely due to chance or exhibits a statistical dependency, measures are used that compute, to what extent the co-occurrence of A and B is statistically significant.
- Significance formulas involve
  - n: number of 'experiments'
  - n_A: number of experiments where A takes part
  - n_B: number of experiments where B takes part
  - n_{AB}: number of experiments where A and B co-occur

EXAMPLES OF CO-OCCURRENCE GRAPHS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Exit Music (for a film) - Radiohead

Wake from your sleep
The drying of your tears
Today we escape
We escape.
Pack and get dressed
Before your father hears us
Before all hell breaks loose.
Breathe keep breathing
Don't lose your nerve.
Breathe keep breathing
I can't do this alone.
Sing us a song
A song to keep us warm
There is such a chill
Such a chill.
You can laugh
A spineless laugh
We hope that your rules
and wisdom choke you
Now we are one
In everlasting peace
We hope that you choke,
that you choke
We hope that you choke,
that you choke
We hope that you choke,
that you choke

SIGNIFICANCE MEASURES

n: number of 'experiments'
n_A: number of experiments where A takes part
n_B: number of experiments where B takes part
n_{AB}: number of experiments where A and B co-occur

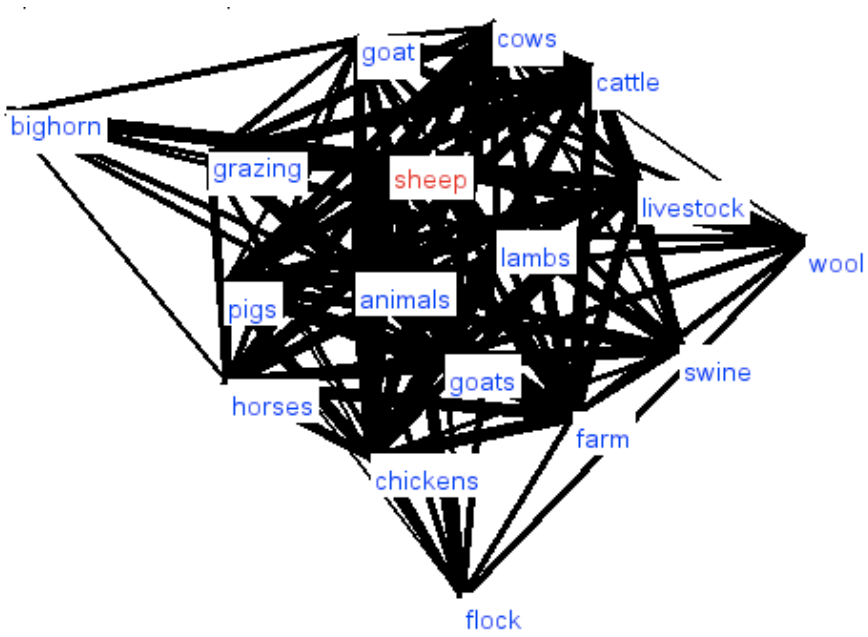
- Joint Frequency sig(A,B) = n_{AB}
- Pointwise Mutual Information $$MI(A,B) = \log_2 \frac{n_{AB}}{n_A \cdot n_B}$$
- Lexicographer's Pointwise Mutual Information $$LMI(A,B) = n_{AB} \cdot \log_2 \frac{n_{AB}}{n_A \cdot n_B}$$
- Log-Likelihood ratio $$-2\log\lambda = 2 \cdot \log \frac{L(H_1)}{L(H_0)} = 2 \left[\begin{array}{l} n\log n - n_A \log n_A - n_B \log n_B + n_{AB} \log n_{AB} \\ + (n - n_A - n_B + n_{AB}) \log (n - n_A - n_B + n_{AB}) \\ + (n_A - n_{AB}) \log (n_A - n_{AB}) \\ + (n_B - n_{AB}) \log (n_B - n_{AB}) \\ - (n - n_A) \log (n - n_A) - (n - n_B) \log (n - n_B) \end{array} \right]$$
- … or any other statistical significance test

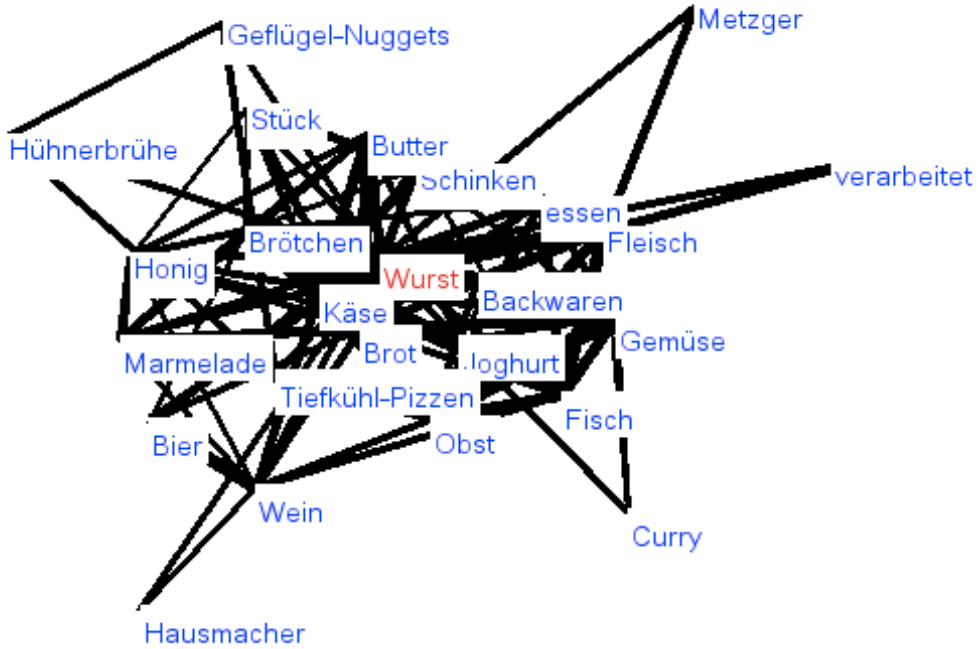
CO-OCCURRENCE GRAPHS FROM LARGE CORPORA

HTTP://CORPORA.INFORMATIK.UNI-LEIPZIG.DE/

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Local neighborhoods: top significant co-occurrences, and their top significant connections
- Mix of syntactic and semantic relations

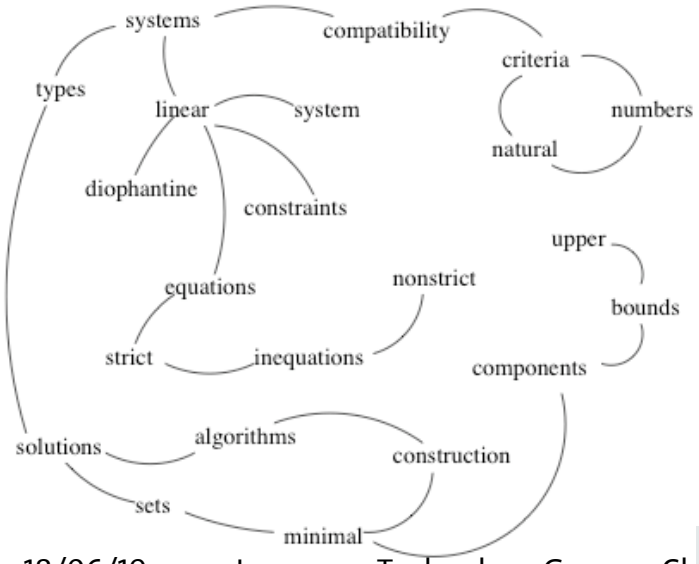




TEXTRANK FOR KEYWORD EXTRACTION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Keyword extraction: find the most salient keywords for a document
- Keyword extraction with PageRank:
  - preprocess document: identify adjectives and nouns as targets
  - target co-occurrence graph: targets co-occurring within a window of 2-10 words
  - apply PageRank to get ranking scores on nodes
  - select highest scoring keywords, possibly concatenate ADJ-NOUN-NOUN sequences if present in the text



Keywords assigned by TextRank:
linear constraints; linear diophantine equations; natural numbers; nonstrict inequations; strict inequations; upper bounds

Keywords assigned by human annotators:
linear constraints; linear diophantine equations; minimal generating sets; non–strict inequations; set of natural numbers; strict inequations; upper bounds

KEYWORD EXTRACTION EVALUATION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

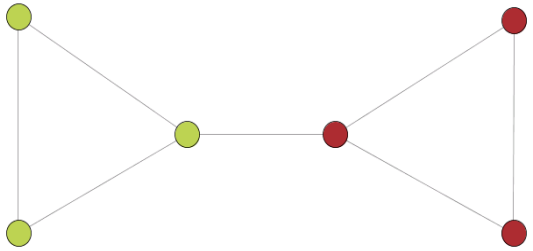
Method	Assigned		Correct		Precision	Recall	F-measure
	Total	Mean	Total	Mean			
TextRank							
Undirected, Co-occ.window=2	6,784	13.7	2,116	4.2	**31.2**	43.1	**36.2**
Undirected, Co-occ.window=3	6,715	13.4	1,897	3.8	28.2	38.6	32.6
Undirected, Co-occ.window=5	6,558	13.1	1,851	3.7	28.2	37.7	32.2
Undirected, Co-occ.window=10	6,570	13.1	1,846	3.7	28.1	37.6	32.2
Directed, forward, Co-occ.window=2	6,662	13.3	2,081	4.1	31.2	42.3	35.9
Directed, backward, Co-occ.window=2	6,636	13.3	2,082	4.1	31.2	42.3	35.9
Hulth (2003)							
Ngram with tag	7,815	15.6	1,973	3.9	25.2	**51.7**	33.9
NP-chunks with tag	4,788	9.6	1,421	2.8	29.7	37.2	33.0
Pattern with tag	7,012	14.0	1,523	3.1	21.7	39.9	28.1

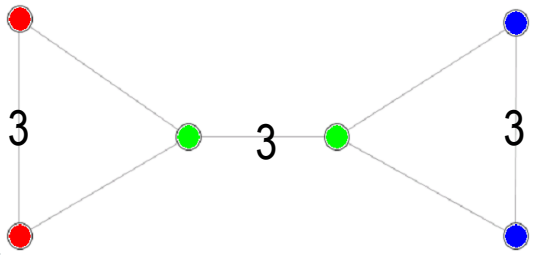
- Comparison: Supervised system that is trained on manually assigned keywords, using frequency and contextual features
- Note that TextRank is unsupervised: no training necessary

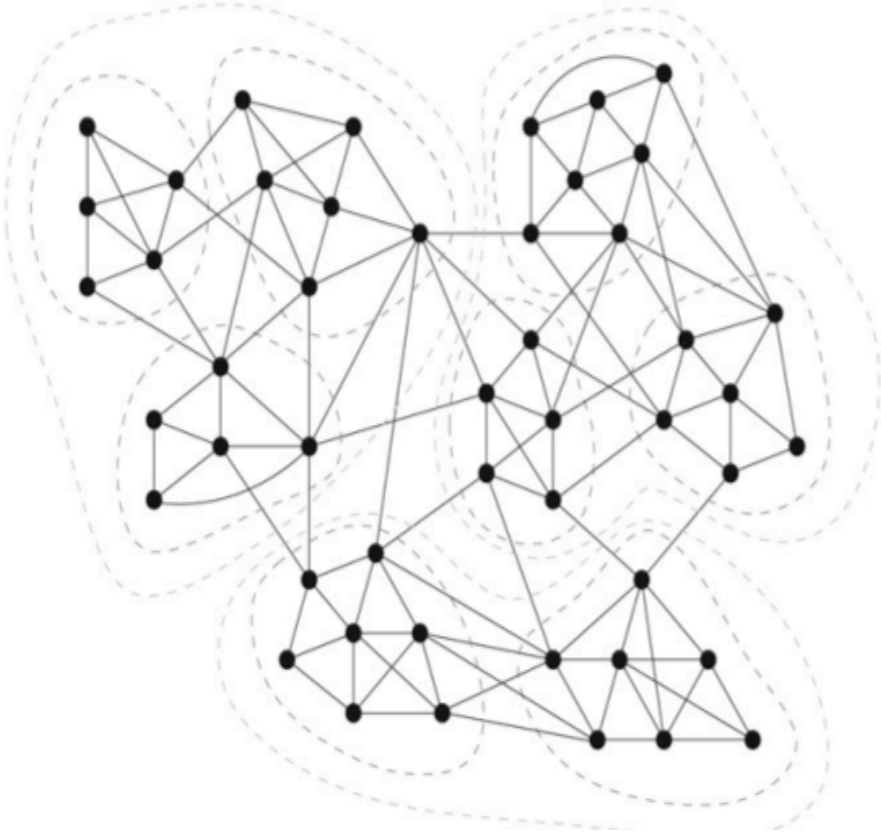
GRAPH CLUSTERING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

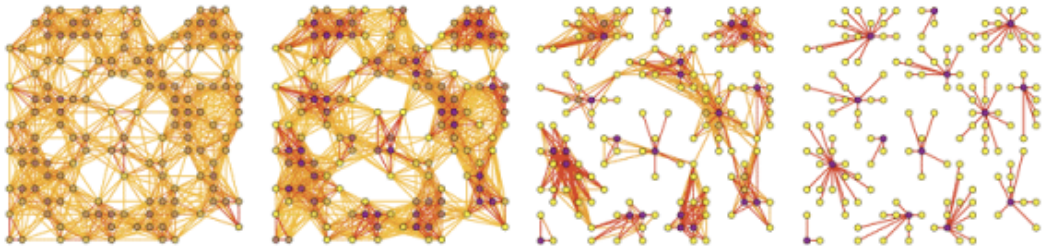
- Task: Find meaningful groups of nodes in graph by cutting edges
- Intuition: Connectedness within a cluster is higher than between clusters
- Many graph clustering algorithms detect the number of clusters







MARKOV CHAIN CLUSTERING



- Clustering based on random walks: MCL is the parallel simulation of all possible random walks up to a finite length on a graph G
- Idea: a random walker on the graph is more likely to stay within the same cluster than to end up in a different cluster after a small number of steps
- Algorithm: can show convergence to a limit T

```
Add loops: transition matrix T= column-normalize (A_G + I)
MCL process: alternate between
   T=T^t            // expansion: raise T to its power of t
   T=inflate(T)     // inflation: increase contrast within
                       columns by raising values to their power
                          of s (s>0) and normalize column-wise
Interpret T as a clustering: use strongest connection as label
```

Stijn van Dongen, Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, May 2000.

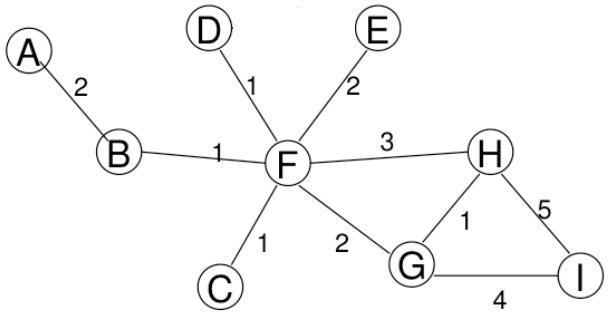
BAD EXPANSION STEP: SIMULATE THE RANDOM WALK

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- (stochastic) adjacency matrix T: probabilities to walk from node in column to node in row in a single step.
- T^2: probabilities to walk from A to B in 2 steps.

	A	B	C	D	E	F	G	H	I
A	***1***	2							
B	2	***1***				1			
C			***1***			1			
D				***1***		1			
E					***1***	2			
F		1	1	1	2	***1***	2	3	
G						2	***1***	1	4
H						3	1	***1***	5
I							4	5	***1***

A_G *loops* added



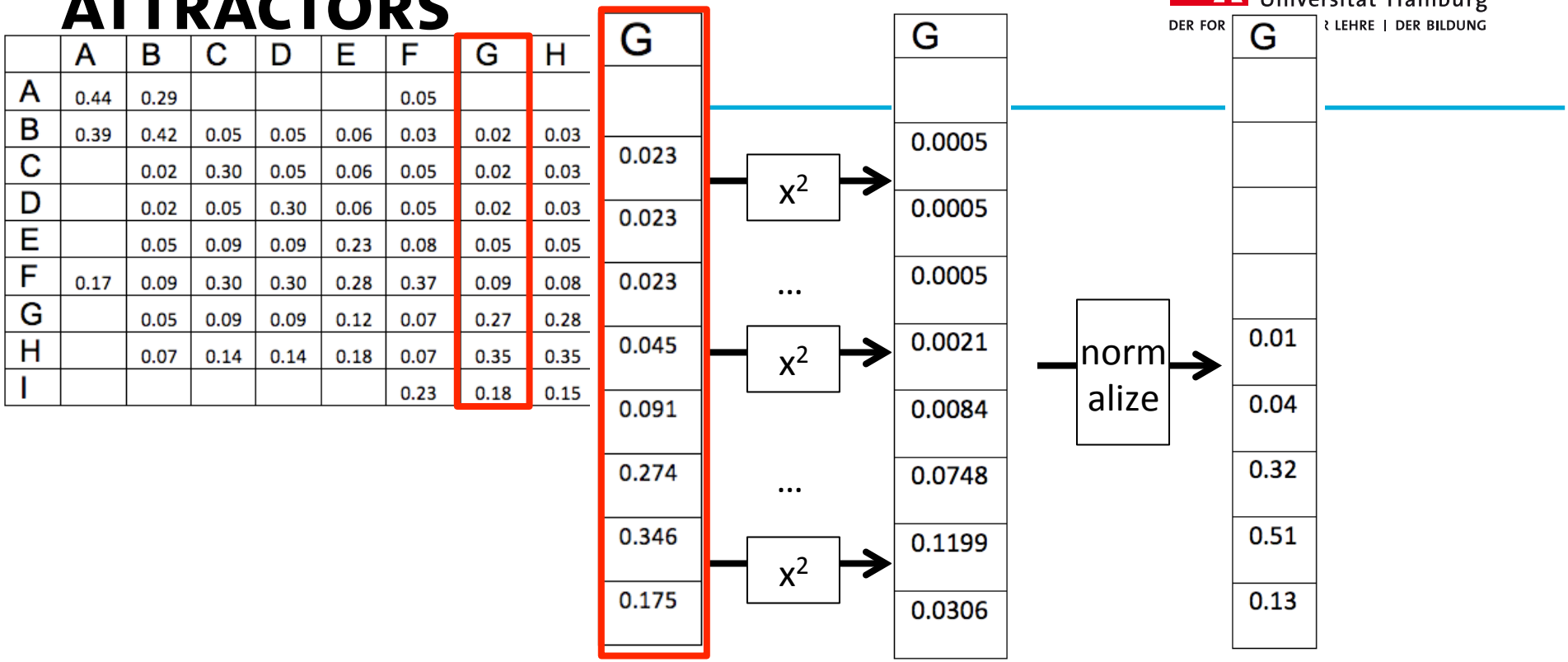
	A	B	C	D	E	F	G	H	I
A	1/3	1/2							
B	2/3	1/4				1/11			
C			1/2			1/11			
D				1/2		1/11			
E					1/3	2/11			
F		1/4	1/2	1/2	2/3	1/11	1/4	3/10	
G						2/11	1/8	1/10	2/5
H						3/11	1/8	1/10	1/2
I							1/2	1/2	1/10

T

	A	B	C	D	E	F	G	H	I
A	0.44	0.29				0.05			
B	0.39	0.42	0.05	0.05	0.06	0.03	0.02	0.03	
C		0.02	0.30	0.05	0.06	0.05	0.02	0.03	
D		0.02	0.05	0.30	0.06	0.05	0.02	0.03	
E		0.05	0.09	0.09	0.23	0.08	0.05	0.05	
F	0.17	0.09	0.30	0.30	0.28	0.37	0.09	0.08	0.25
G		0.05	0.09	0.09	0.12	0.07	0.27	0.28	0.14
H		0.07	0.14	0.14	0.18	0.07	0.35	0.35	0.15
I						0.23	0.18	0.15	0.46

T^2

BAD INFLATION STEP: ONLY KEEP ATTRACTORS



	A	B	C	D	E	F	G	H
A	0.44	0.29				0.05		
B	0.39	0.42	0.05	0.05	0.06	0.03	0.02	0.03
C		0.02	0.30	0.05	0.06	0.05	0.02	0.03
D		0.02	0.05	0.30	0.06	0.05	0.02	0.03
E		0.05	0.09	0.09	0.23	0.08	0.05	0.05
F	0.17	0.09	0.30	0.30	0.28	0.37	0.09	0.08
G		0.05	0.09	0.09	0.12	0.07	0.27	0.28
H		0.07	0.14	0.14	0.18	0.07	0.35	0.35
I						0.23	0.18	0.15

- Inflate the differences within a column by taking the k-th power of the value, then normalize to ensure stochastic property. k regulates the cluster sizes
- Clustering: Highest entry in column vector is cluster label

Variants:

- Could add small random noise to break ties
- Optimization: Only keep K largest values, only keep values over threshold

CHINESE WHISPERS GRAPH CLUSTERING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

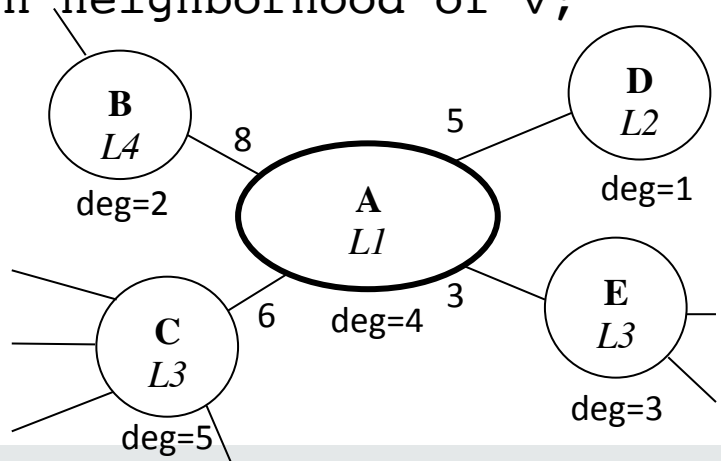
- MCL: keep only a few strong neighbors
- Chinese Whispers: only propagate strongest label in neighborhood

```
initialize:
    forall v_i in V: class(v_i)=i;

while changes:

  forall v in V, randomized order:

   class(v)=highest ranked
       class in neighborhood of v;
```



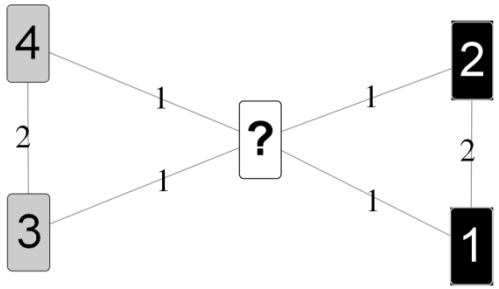
- Nodes have a class and communicate it to their adjacent nodes
- A node adopts one of the majority class in its neighbourhood
- Nodes are processed in random order for some iterations
- Node weighting schemes

PROPERTIES OF CHINESE WHISPERS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Advantages:
  - Efficiency: CW is time-linear in the number of edges. This is bound by n^2 with n= number of nodes, but in real world data, graphs are much sparser
  - Parameter-free: this includes number of clusters

- Disadvantages:
  - Non-deterministic: due to random order processing and possible ties w.r.t. the majority.
  - Does not converge: See bowtie tie example:



The disadvantages are usually not severe for NL data.

Biemann, C. (2006): Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. Proceedings of the HLT-NAACL-06 Workshop on Textgraphs-06, New York, USA.

APPLICATION: LANGUAGE SEPARATION

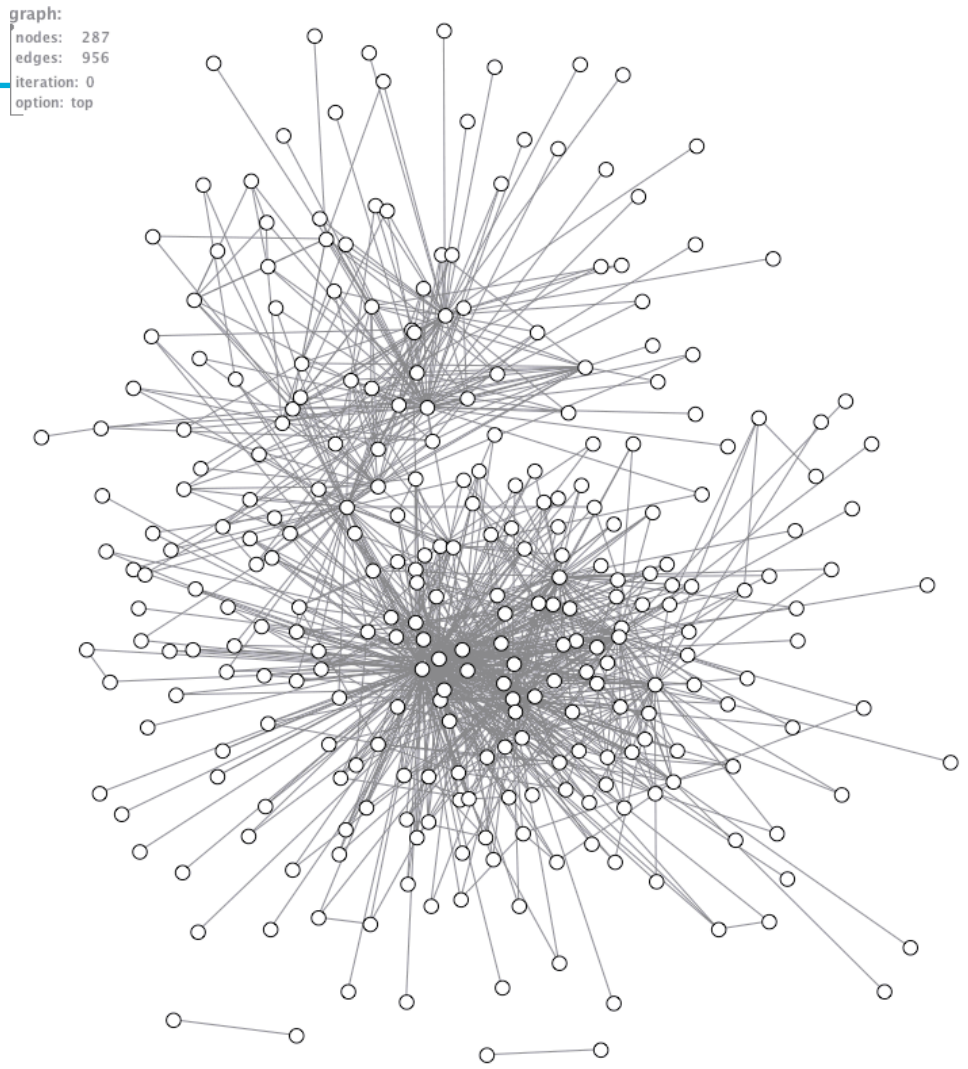
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- When downloading the internet, we don't know what languages are there. It would be useful to sort the pages by language
- Idea: Train a language classifier on text of known language
- Material for supervised training:
  - German: Als ich in der Schule war, gab es noch keine Handys. Es war noch nicht möglich, in der Pause zu telefonieren.
  - English: It does not rain a lot in the Sahara desert. In a way, it is one of the driest regions of the world.
- Model: most frequent words:
  - German: ich, war, es, noch, in
  - English: it, the, of, in, a
- Classification:
  - This is a test. It tests a language identification system.
  - Es ist noch kein Meister in den Brunnen gefallen.
  - Soldiers die in the war.
  - Mi aerodeslizador está lleno de anguilas

UNSUPERVISED LANGUAGE SEPARATION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

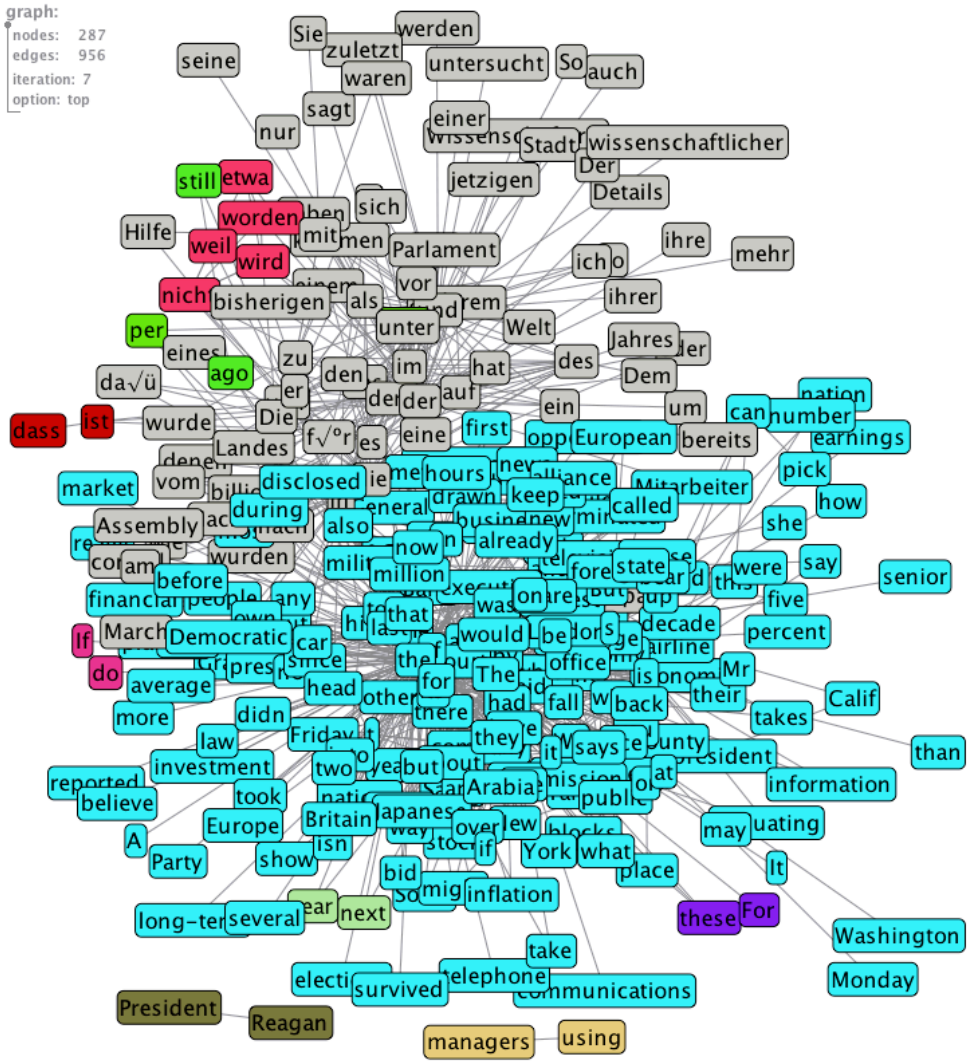
- Multilingual word co-occurrence graph is connected: some words are used in several languages, the same names are also used throughout.
- Cluster the multilingual word co-occurrence graph and use cluster members to identify sentences/documents belonging to that cluster



CLUSTERING WITH CHINESE WHISPERS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Random Initialization
- Iteration 1
- Iteration 2
- Iteration 3
- Iteration 4
- Iteration 5
- Stop



DISTRIBUTIONAL HYPOTHESIS (RECAP)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

The **Distributional Hypothesis** in linguistics is the theory that words that occur in similar contexts tend to have similar meanings.

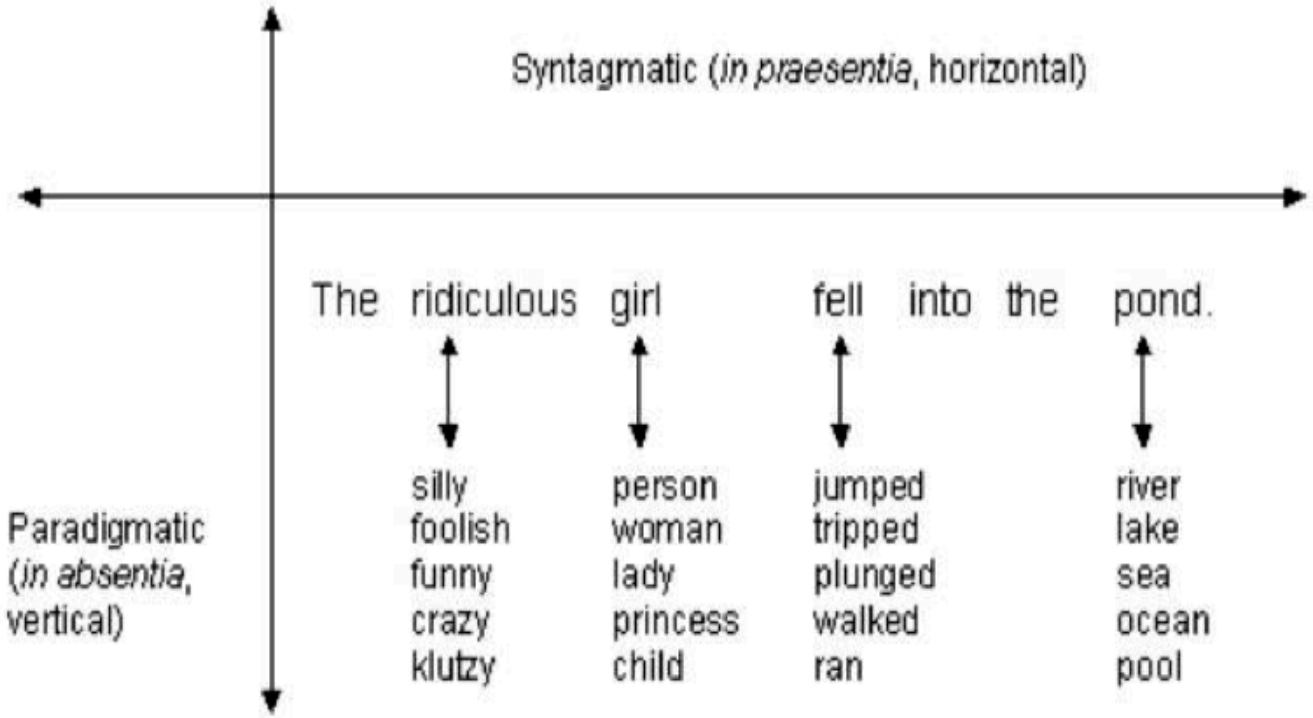
The Distributional Hypothesis is the basis for **Statistical Semantics**. It states that the meaning of a word can be defined in terms of its context.

Distributional Hypothesis:

- words are characterized by their (typical) contexts: meaning of a word can be defined in terms of its context
- words are more similar, the more contexts they share

Any process that builds a structure on sentences can be used as a source for contexts

SYNTAGMATIC VS. PARADIGMATIC RELATIONS



http://courses.nus.edu.sg/course/elltankw/history/Vocab/B.htm

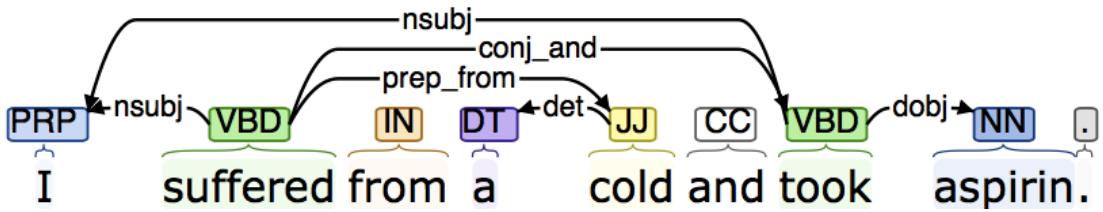
- Syntagmatic Relations: syntactic constraints in the context
- Paradigmatic Relations: associations, semantic constraints

Ferdinand de Saussure. Cours de linguistique generale. Paris: Payot, 1916

THE @@ OPERATION: PRODUCING PAIRS OF TERMS AND CONTEXTS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SENTENCE:



STANFORD COLLAPSED DEPENDENCIES: *http://nlp.stanford.edu:8080/parser/*

nsubj(suffered, I); nsubj(took, I); root(ROOT, suffered); det(cold, a); prep_from(suffered, cold); conj_and(suffered, took); dobj(took, aspirin)

TERM-CONTEXT PAIRS:

suffered	nsubj(@@, I)	1
took	nsubj(@@, I)	1
cold	det(@@, a)	1
suffered	prep_from(@@, cold)	1
suffered	conj_and(@@, took)	1
took	dobj(@@, aspirin)	1

I	nsubj(suffered, @@)	1
I	nsubj(took, @@)	1
a	det(cold, @@)	1
cold	prep_from(suffered, @@)	1
took	conj_and(suffered, @@)	1
aspirin	dobj(took, @@)	1

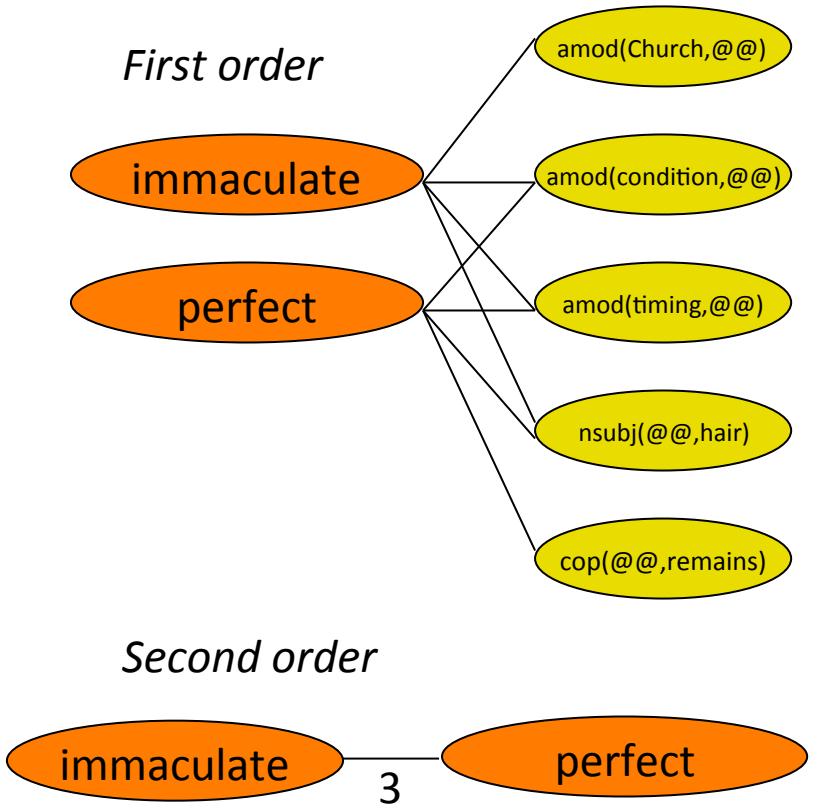
DISTRIBUTIONAL THESAURUS (DT)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Computed from distributional similarity statistics
- **Entry** for a **target** word consists of a ranked list of neighbors

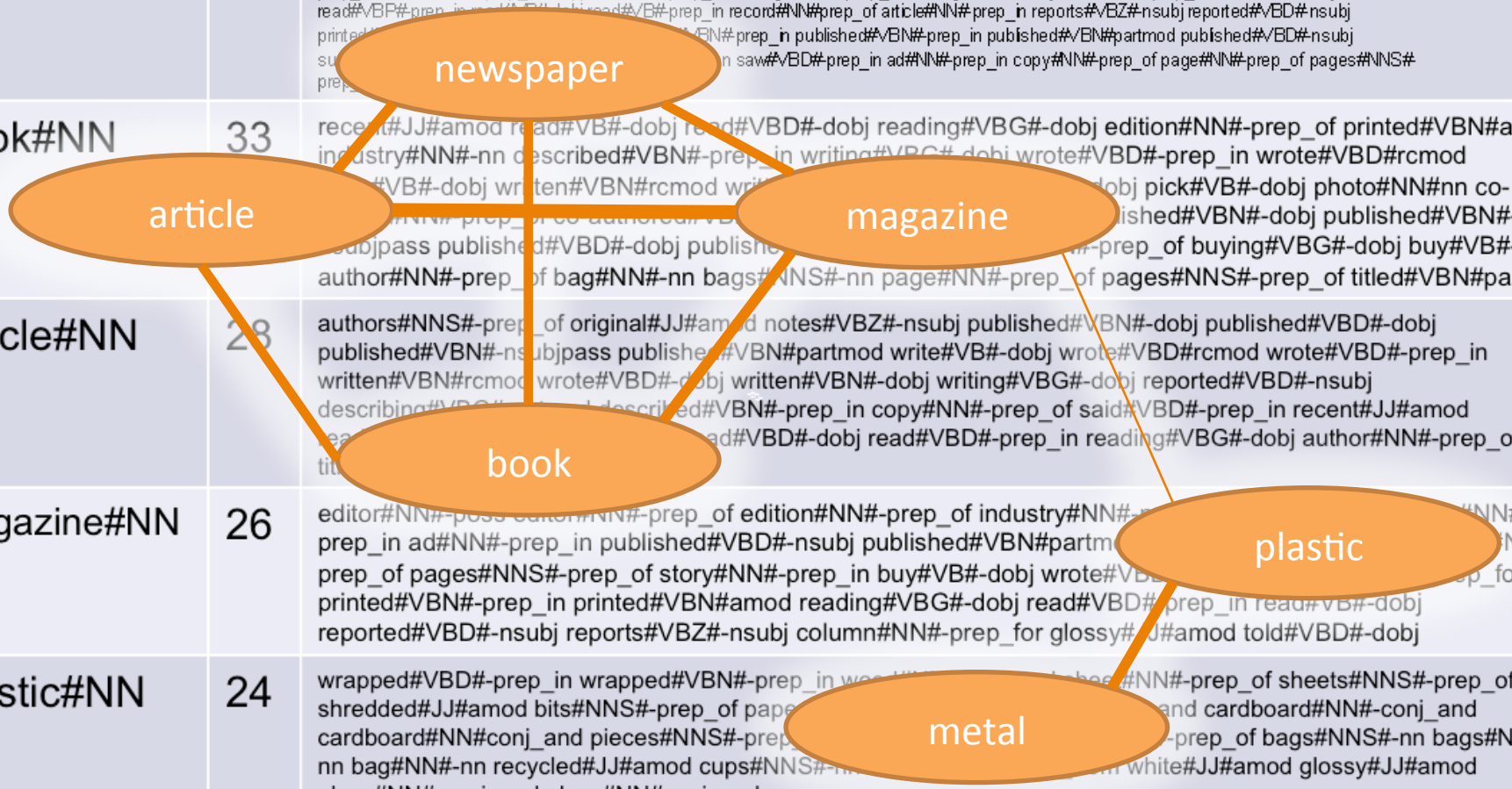
meeting	
meeting	288
meetings	102
hearing	89
session	68
conference	62
summit	51
forum	46
workshop	46
hearings	46
ceremony	45
sessions	41
briefing	40
event	40
convention	38
gathering	36
...	

articulate	
articulate	89
explain	19
understand	17
communicate	17
defend	16
establish	15
deliver	14
evaluate	14
adjust	14
manage	13
speak	13
change	13
answer	13
maintain	13
...	



DT ENTRY "PAPER#NN" WITH CONTEXTS

paper#NN	s	common contexts
newspaper#NN	45	told#VBD#-dobj column#NN#-prep_in local#JJ#amod editor#NN#-poss edition#NN#-prep_of editor#NN#-prep_of hometown#NN#nn industry#NN#-nn clips#NNS#-nn shredded#JJ#amod pick#VB#-dobj news#NNP#appos daily#JJ#amod writes#VBZ#-nsubj write#VB#-prep_for wrote#VBD#-prep_for wrote#VBD#-prep_in wrapped#VBN#-prep_in reading#VBG#-prep_in reading#VBG#-dobj read#VBD#-prep_in read#VBD#-dobj read#VBP#-prep_in record#NN#prep_of article#NN#-prep_in reports#VBZ#-nsubj reported#VBD#-nsubj printed#VBN#-prep_in published#VBN#-prep_in published#VBN#partmod published#VBD#-nsubj saw#VBD#-prep_in ad#NN#-prep_in copy#NN#-prep_of page#NN#-prep_of pages#NNS#
book#NN	33	recent#JJ#amod read#VB#-dobj read#VBD#-dobj reading#VBG#-dobj edition#NN#-prep_of printed#VBN#amod industry#NN#-nn described#VBN#-prep_in writing#VBG#-dobj wrote#VBD#-prep_in wrote#VBD#rcmod #VB#-dobj written#VBN#rcmod dobj pick#VB#-dobj photo#NN#nn co- ished#VBN#-dobj published#VBN#- ubjpass published#VBD#-dobj publish -prep_of buying#VBG#-dobj buy#VB#-dobj author#NN#-prep_of bag#NN#-nn bags#NNS#-nn page#NN#-prep_of pages#NNS#-prep_of titled#VBN#partmod
article#NN	28	authors#NNS#-prep_of original#JJ#amod notes#VBZ#-nsubj published#VBN#-dobj published#VBD#-dobj published#VBN#-nsubjpass publish #VBN#partmod write#VB#-dobj wrote#VBD#rcmod wrote#VBD#-prep_in written#VBN#rcmod wrote#VBD#-dobj written#VBN#-dobj writing#VBG#-dobj reported#VBD#-nsubj describing#VBG# described#VBN#-prep_in copy#NN#-prep_of said#VBD#-prep_in recent#JJ#amod ad#VBD#-dobj read#VBD#-prep_in reading#VBG#-dobj author#NN#-prep_of
magazine#NN	26	editor#NN#-poss edition#NN#-prep_of edition#NN#-prep_of industry#NN#- NN#- prep_in ad#NN#-prep_in published#VBD#-nsubj published#VBN#partm NN#- prep_of pages#NNS#-prep_of story#NN#-prep_in buy#VB#-dobj wrote# p_for printed#VBN#-prep_in printed#VBN#amod reading#VBG#-dobj read#VBD# prep_in read#VB#-dobj reported#VBD#-nsubj reports#VBZ#-nsubj column#NN#-prep_for glossy# J#amod told#VBD#-dobj
plastic#NN	24	wrapped#VBD#-prep_in wrapped#VBN#-prep_in NN#-prep_of sheets#NNS#-prep_of shredded#JJ#amod bits#NNS#-prep_of pape and cardboard#NN#-conj_and cardboard#NN#conj_and pieces#NNS#-prep -prep_of bags#NNS#-nn bags#NN#-nn bag#NN#-nn recycled#JJ#amod cups#NNS# white#JJ#amod glossy#JJ#amod glass#NN#-conj_and glass#NN#conj_and
metal#NN	23	bits#NNS#-prep_of made#VBN#-prep_from work#NN#-nn wood#NN#conj_and scrap#NN#nn paper#NN#-conj_and piece#NN#-prep_of pile#NN#-prep_of pieces#NNS#-prep_of plastic#NN#conj_and plastic#NN#-conj_and plastic#NN#conj_or plate#NN#-nn plates#NNS#-nn recycled#JJ#amod clip#NN#-nn products#NNS#-nn put#VBD#-prep_to put#VB#-prep_to glass#NN#-conj_and glass#NN#conj_and tons#NNS#-prep_of white#JJ#amod

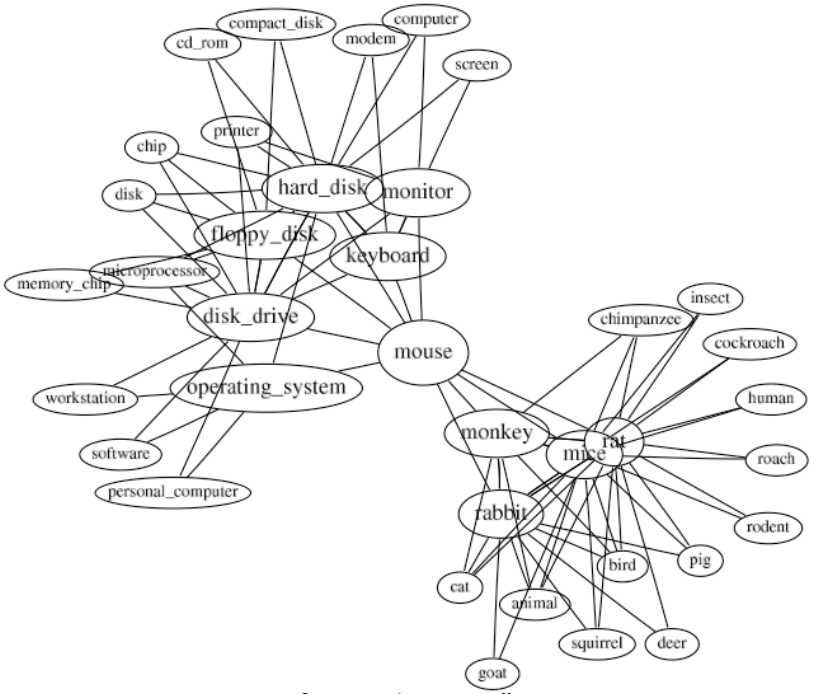


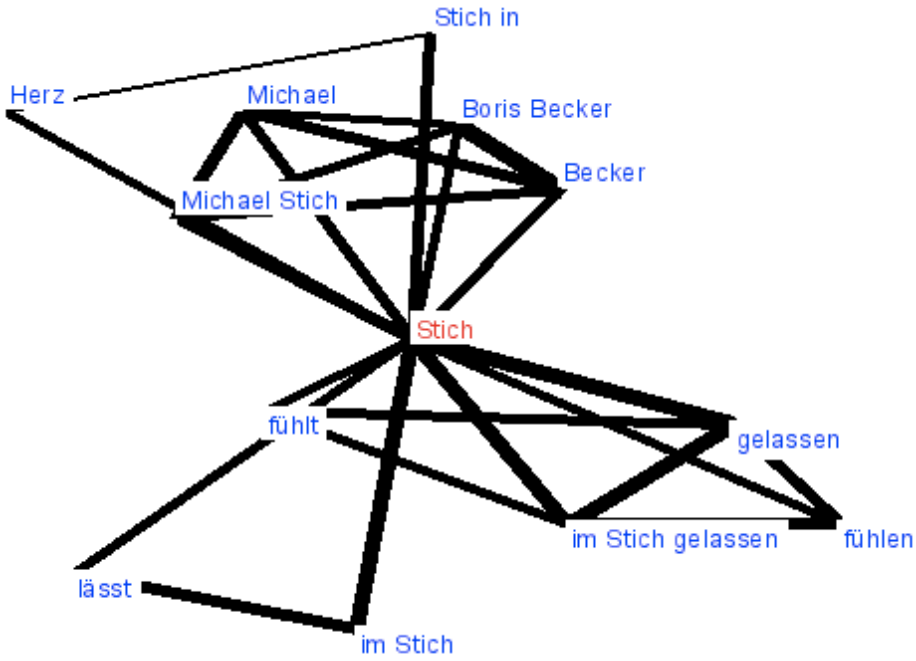
WORD SENSE INDUCTION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Some words have several meanings
- different meanings appear in different contexts
- global contexts are represented by co-occurrences

➔ Cluster the co-occurrence graph to get contexts per meaning



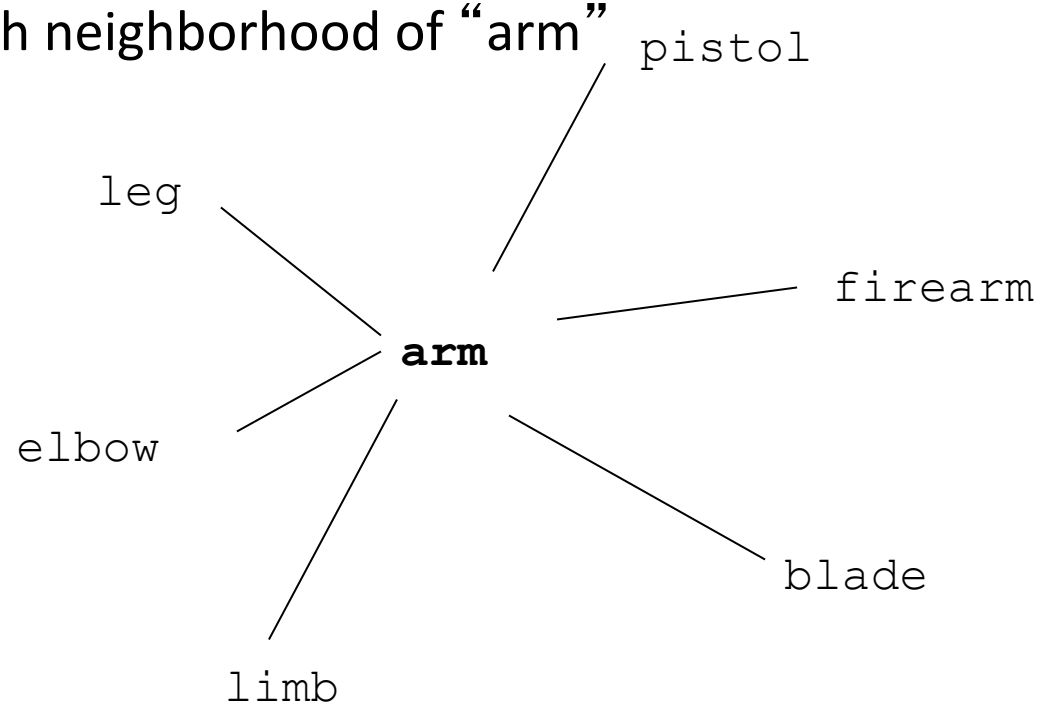


"Discovering corpus-specific word-senses", Beate Dorow, Dominic Widdows, 10th Conference of the European Chapter of the Association for Computational Linguistics (Conference Companion), 2003, pp. 79-82.

AMBIGUITY DETECTION EXAMPLE

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

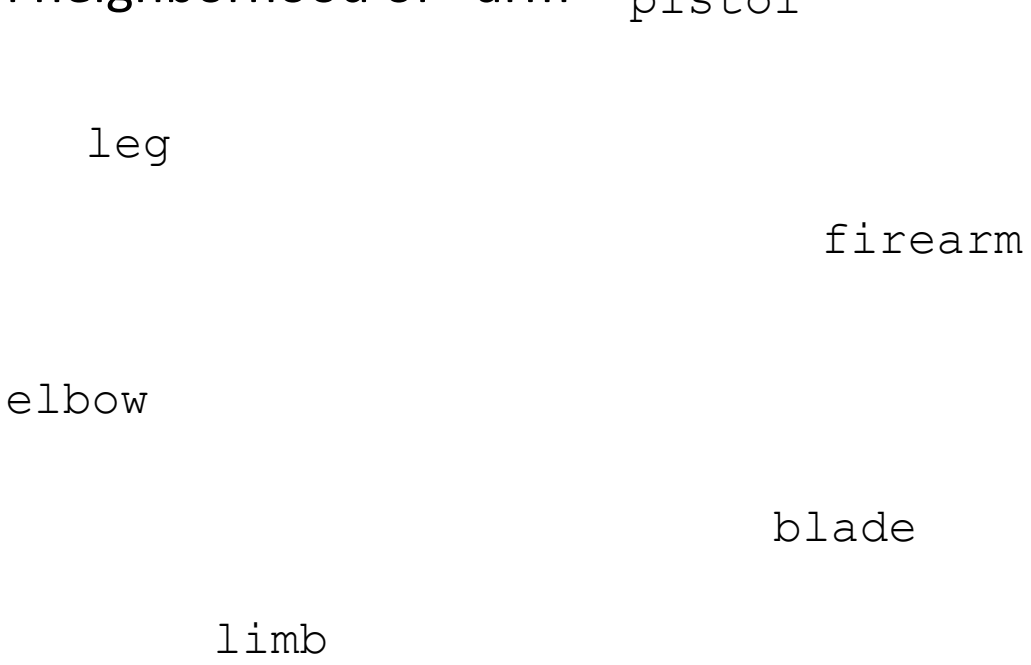
1. Retrieve co-occurrence graph neighborhood of “arm”



STEPS IN CLUSTERING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

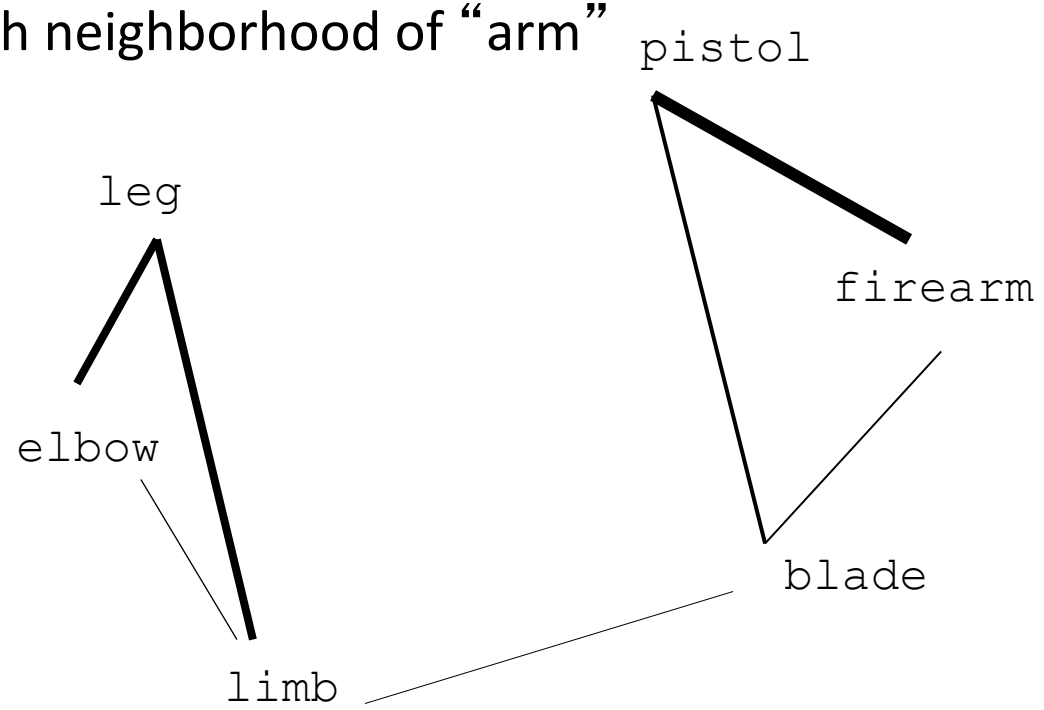
1. Retrieve co-occurrence graph neighborhood of “arm”
2. Remove target word



STEPS IN CLUSTERING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

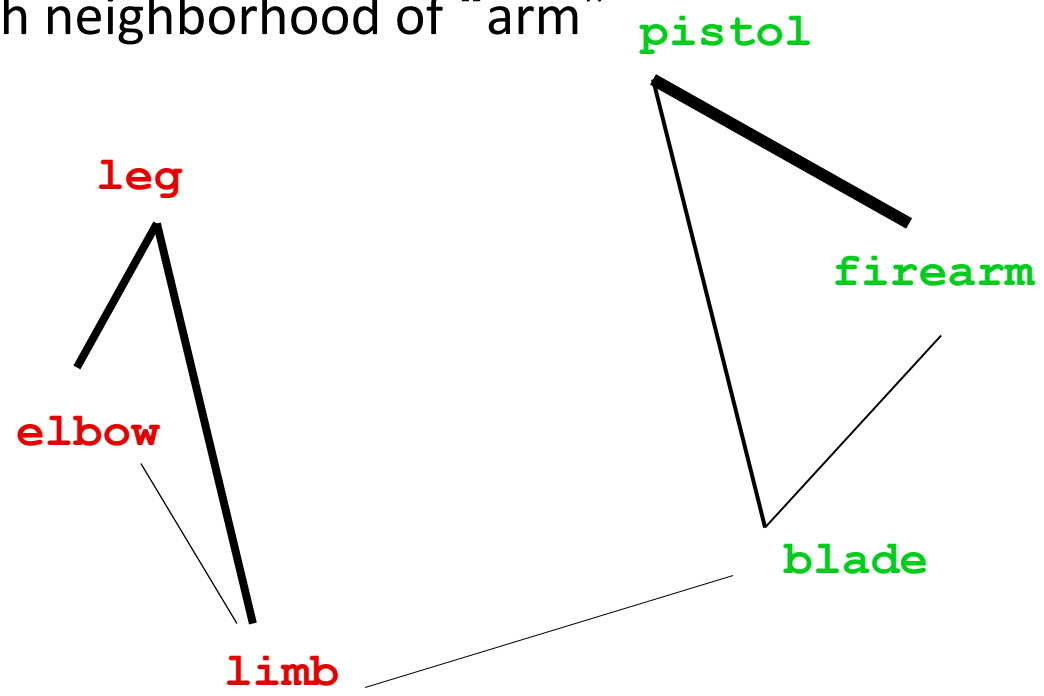
1. Retrieve co-occurrence graph neighborhood of “arm”
2. Remove target word
3. Retrieve edges for present nodes



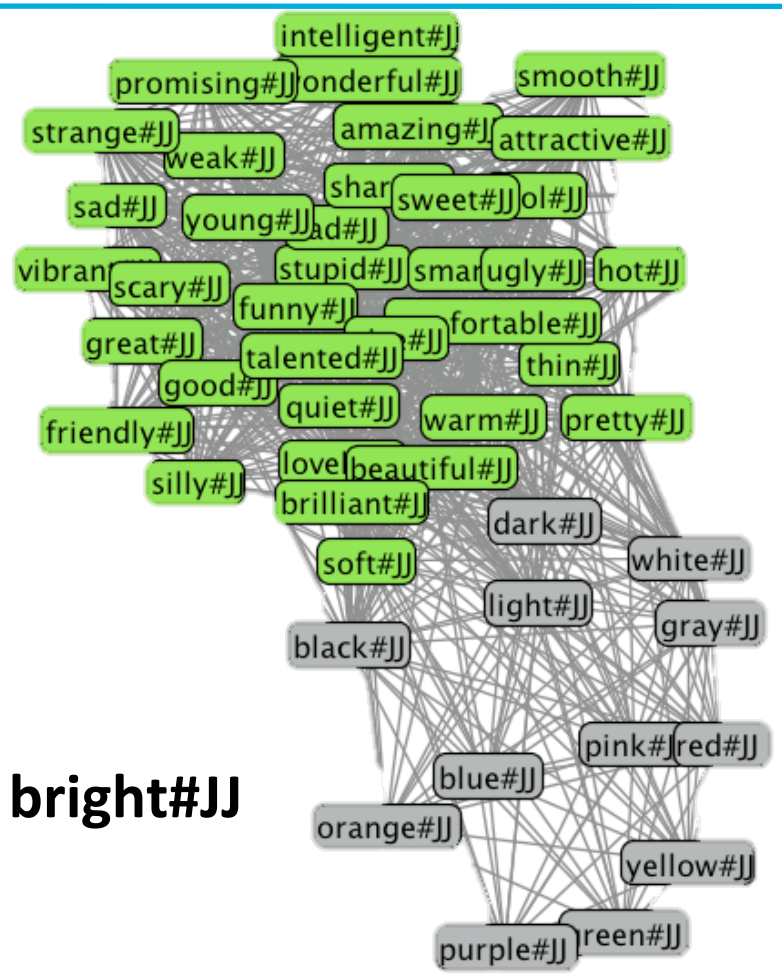
STEPS IN CLUSTERING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

1. Retrieve co-occurrence graph neighborhood of “arm”
2. Remove target word
3. Retrieve edges for present nodes
4. Apply graph clustering

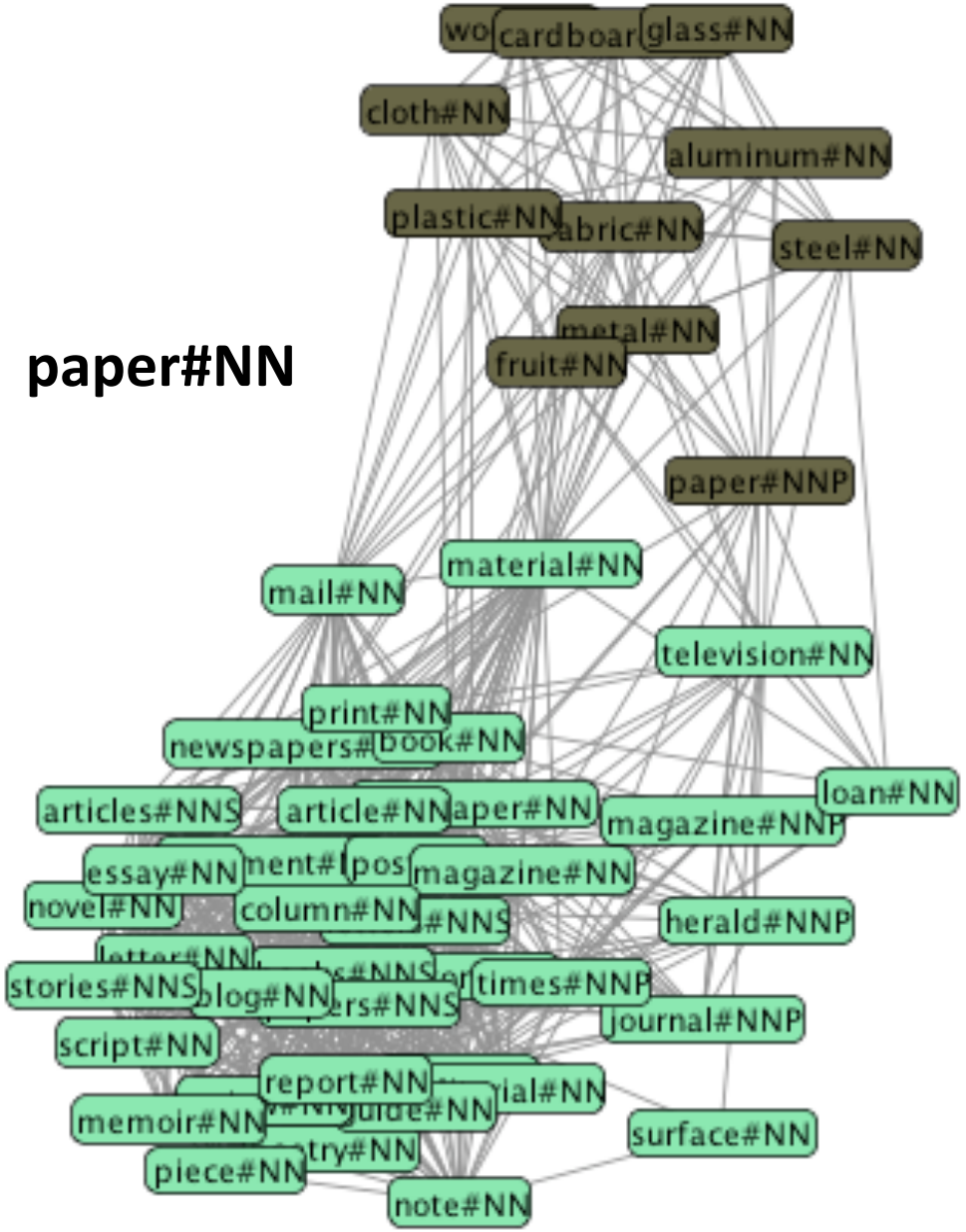


SAMPLE RESULTS



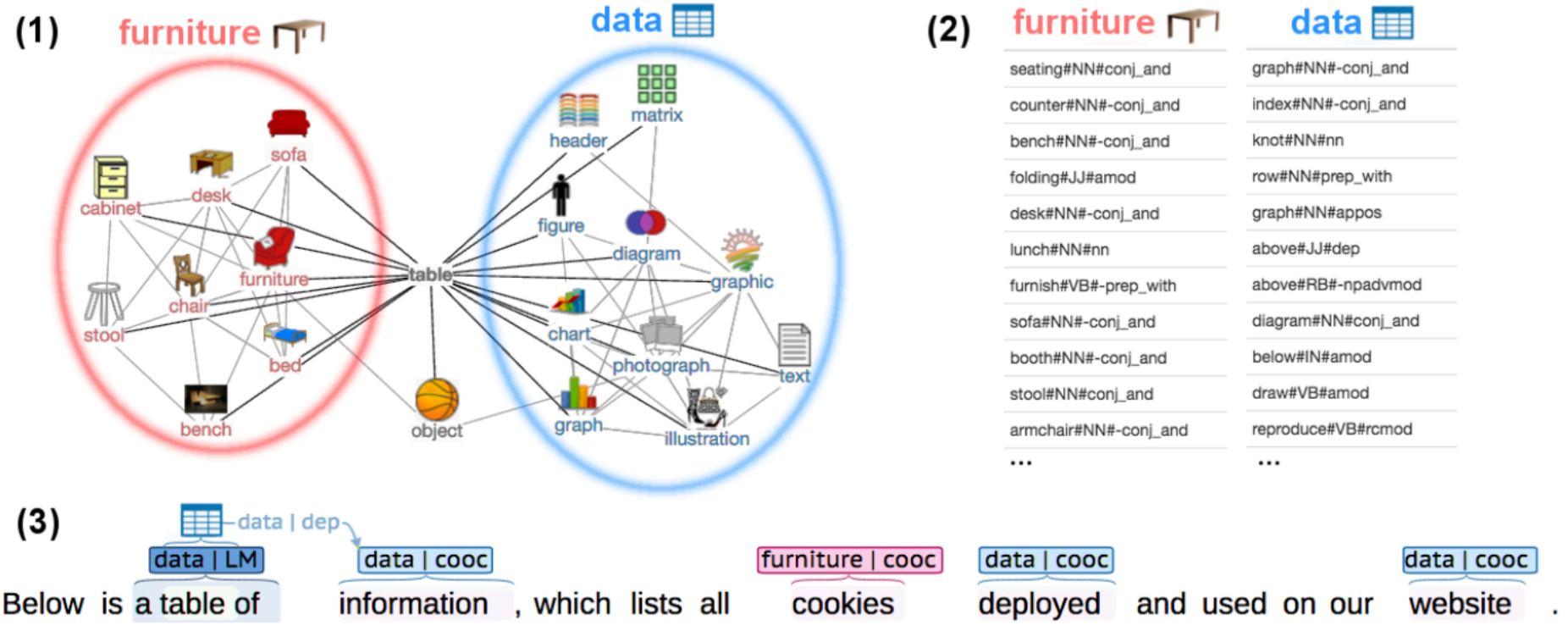
bright#JJ

paper#NN



INTERPRETABLE WORD SENSE INDUCTION AND DISAMBIGUATION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



- Induction of interpretable word sense inventory
- interpretable disambiguation mechanism

Panchenko, A., Ruppert, E., Faralli, S., Ponzetto, S. P. and Biemann, C. (2017): Unsupervised Does Not Mean Uninterpretable: The Case for Word Sense Induction and Disambiguation. Proceedings of EACL 2017, Valencia, Spain.

INTERPRETABILITY AND ROBUSTNESS OF REPRESENTATION

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Why are 'anaconda' and 'python' similar?

largest point of critique on dense vector representations:

- lack of interpretability of dimensions
- when using random sampling methods: re-running the procedure results in different values

because their cosine similarity is 0.95, being most similar in dimensions 54, 3 and 8 while being least similar in dimensions 90, 22 and 15 using random seed 0.

Sparse (graph-based) models:

- readable
- deterministic / reproducible on same corpus
- robust: similar representations on similar corpora

because they share 36 significant syntactic contexts, of which the most salient are:
they coil up, are snakes, swallow, digest, gorge, tighten, and co-occur in conjunctions with other snakes such as rattlesnake, cobra, ..

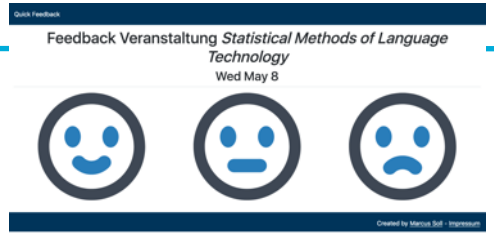
SUMMARY ON GRAPH-BASED METHODS FOR NLP

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Graph representation is a natural representation of items and their relations
- Can use well-known graph algorithms for the solution of specific NLP problems
- Taking the overall structure into account improves some NLP tasks
- Graph clustering algorithms solve unsupervised NLP tasks without the need to specify the number of clusters

IMMEDIATE FEEDBACK

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



knowledge-based, supervised, word sense ductio

WORD SENSE AND MEANING