

JoBimText:

A framework for distributional semantics

LKE Tutorial

02/24/2015

11:00-15:00

By

Martin Riedl and Eugen Ruppert

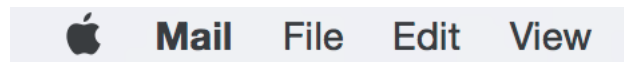


Timetable

- 11:00 – 12:00 Methods and Applications of JoBimText
- 12:00 – 13:00 Using JobimText I
- 13:00 – 14:00 Lunch
- 14:00 – 14:30 Using JoBimText II
- [14:30 – 15:00 Discussion]

Understanding Text

The **bar** serves delicious beer



Click on Mail in the menu **bar**

Use Dictionaries or Ontologies



Photo by zeh fernando under Creative Commons licence

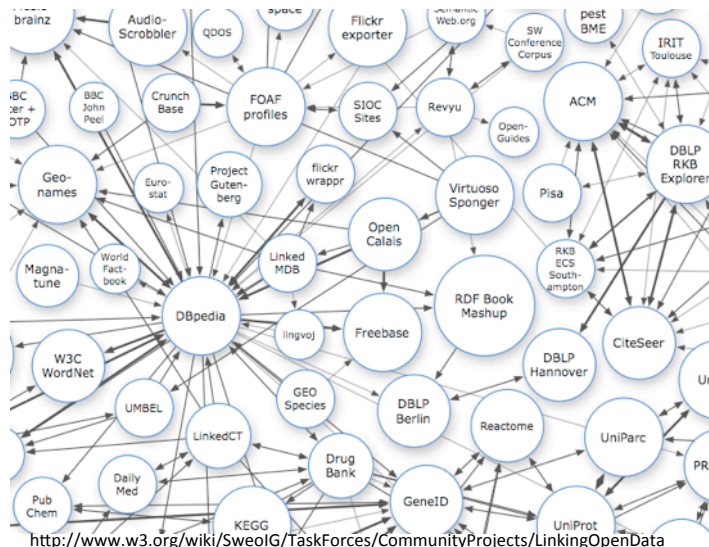
Advantages:

- Sense inventory given
- Linking to concepts
- Full control

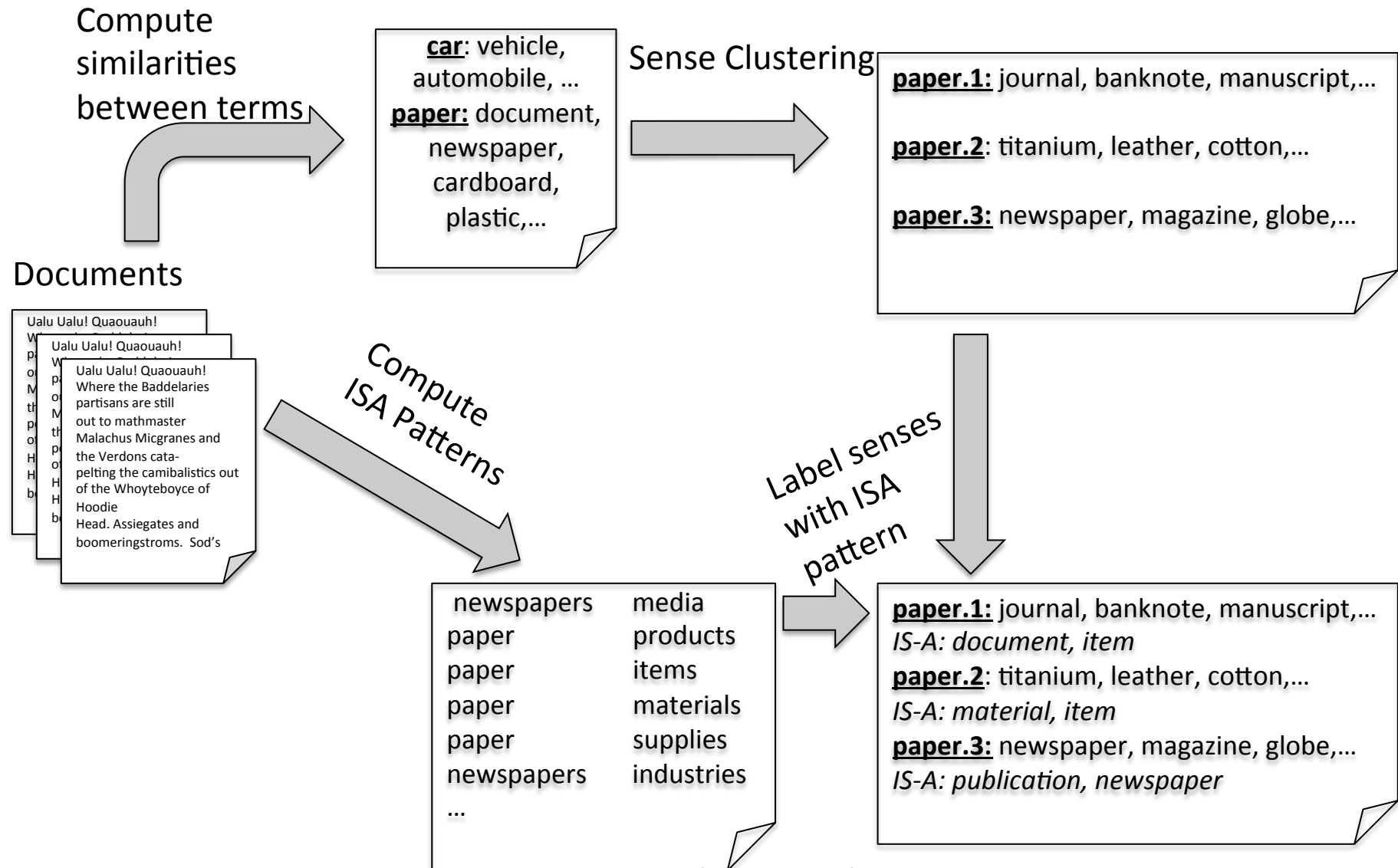
Disadvantages:

- Dictionaries have to be created
- Dictionaries are incomplete
- Language changes constantly:
new words, new meanings ...

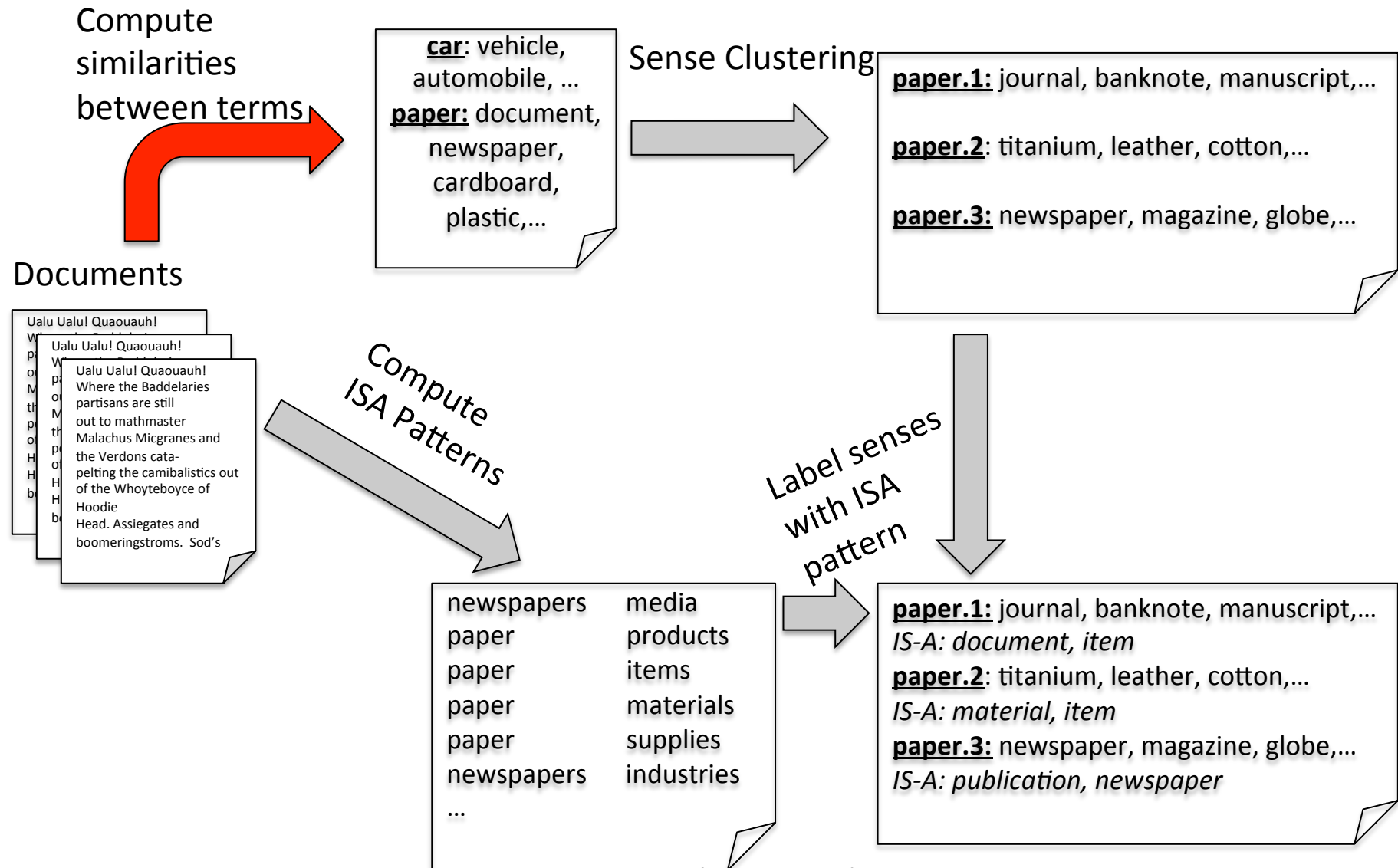
*“give a man a fish and you
feed him for a day...*



The world of JoBimText



The world of JoBimText



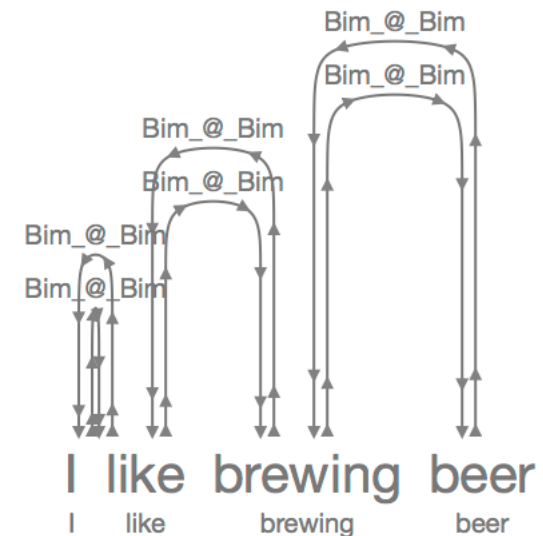
Computing Similarities

- @@-Operation
 - Extract Jo's and Bim's
 - Jo: e.g. word, lemma, ngram, sentence, image, ???
 - Bim: e.g.  neighboring words, dependency parses, words describing image, ???
- Similarity Computation
 - Compute similarities between Jo's [and also Bim's]

Example: Trigram @@-Operation

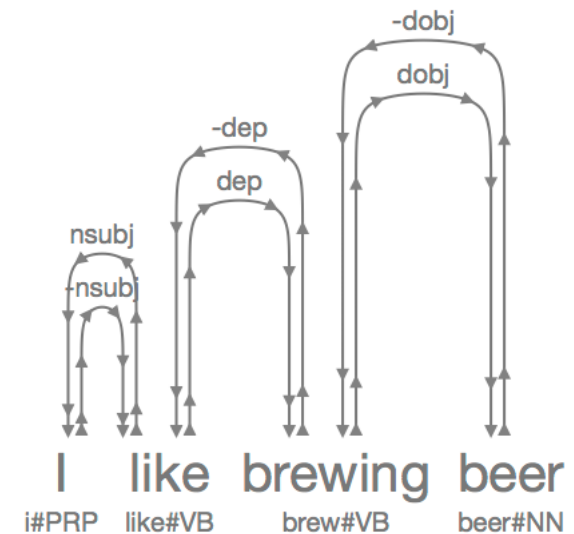
- Input: I like brewing beer.
- Holing Operation:
 - Extract Jo – Bim

Jo	Bim
I	Trigram(_,@,like)
like	Trigram(I,@,brewing)
brewing	Trigram(like,@,beer)
beer	Trigram(brewing,@,_)



Example: Parsing @@-Operation

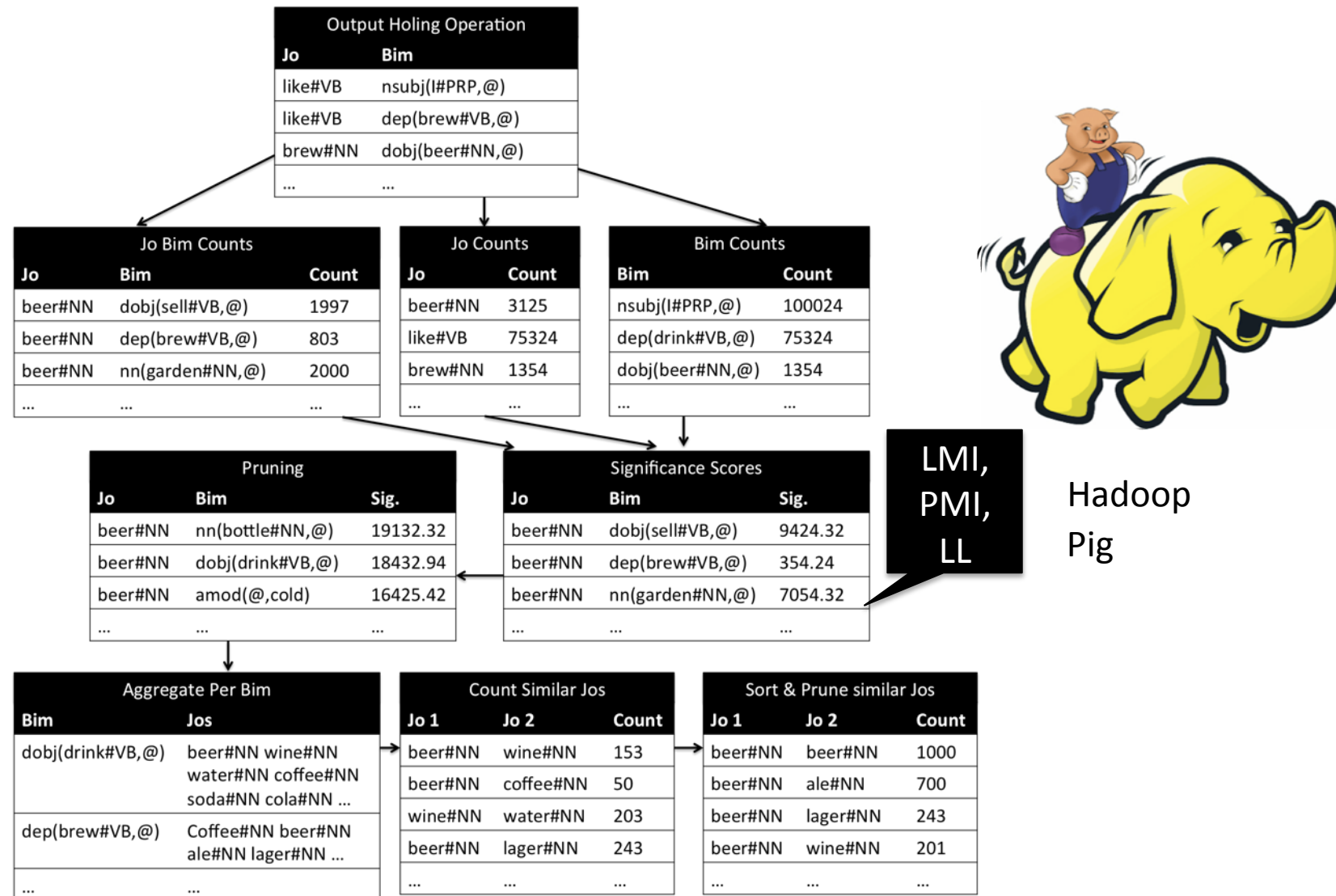
- Input: I like brewing beer.
- Holing Operation:
 - Parsing and Lemmatization
 - Extract Jo – Bim



Jo	Bim
like#VB	nsubj(I#PRP,@)
like#VB	dep(brew#VB,@)
Brew#VB	dobj(beer#NN,@)

Jo	Bim
I#PRP	nsubj(@, like#VB)
brew#VB	dep(@, like#VB)
beer#NN	dobj(@, brew#VB)

Computing Similarities



Distributional Thesaurus based on Jo's

Similar terms for
beer using
dependency
parses on
Newspaper data

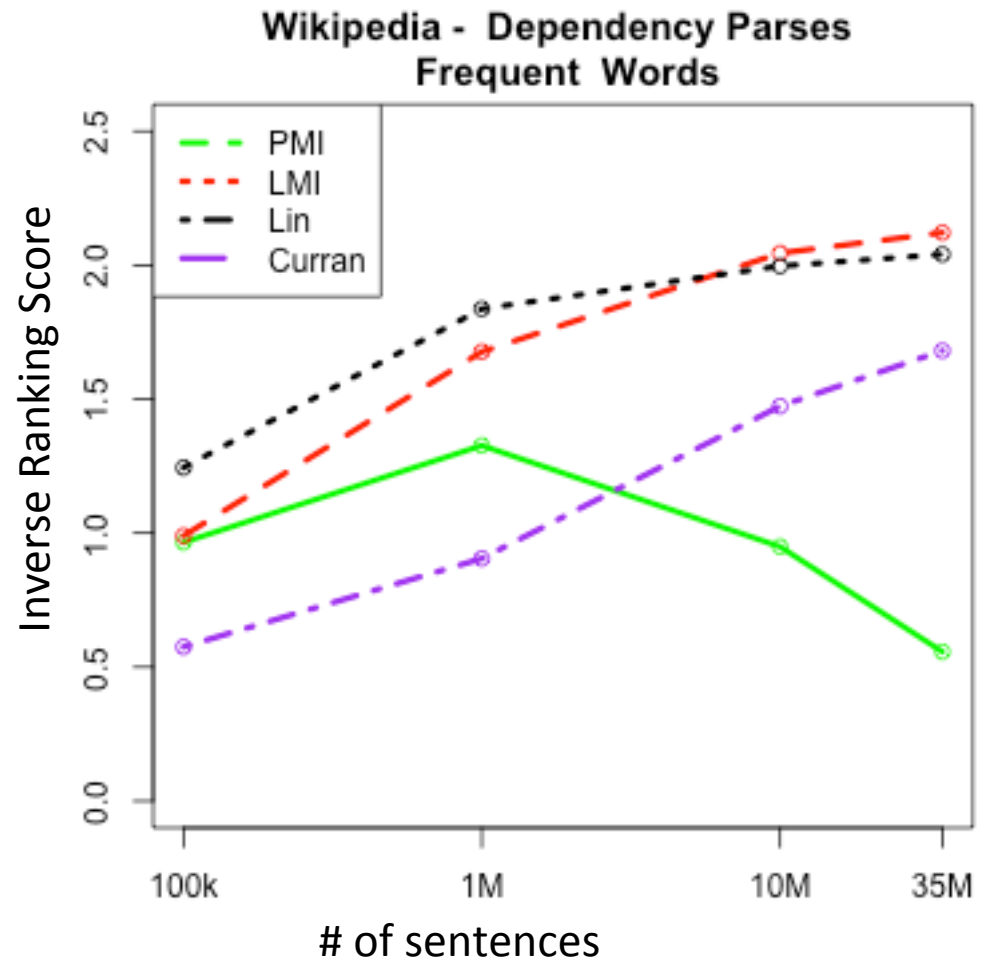
beer#NN	711
drink#NN	190
wine#NN	183
soda#NN	179
coffee#NN	175
beverage#NN	168
liquor#NN	155
tea#NN	154
lager#NN	140
champagne#NN	138
vodka#NN	136
whiskey#NN	126
ale#NN	126
...	...

Similar terms
using n-grams and
trigram holing
based on
medical data

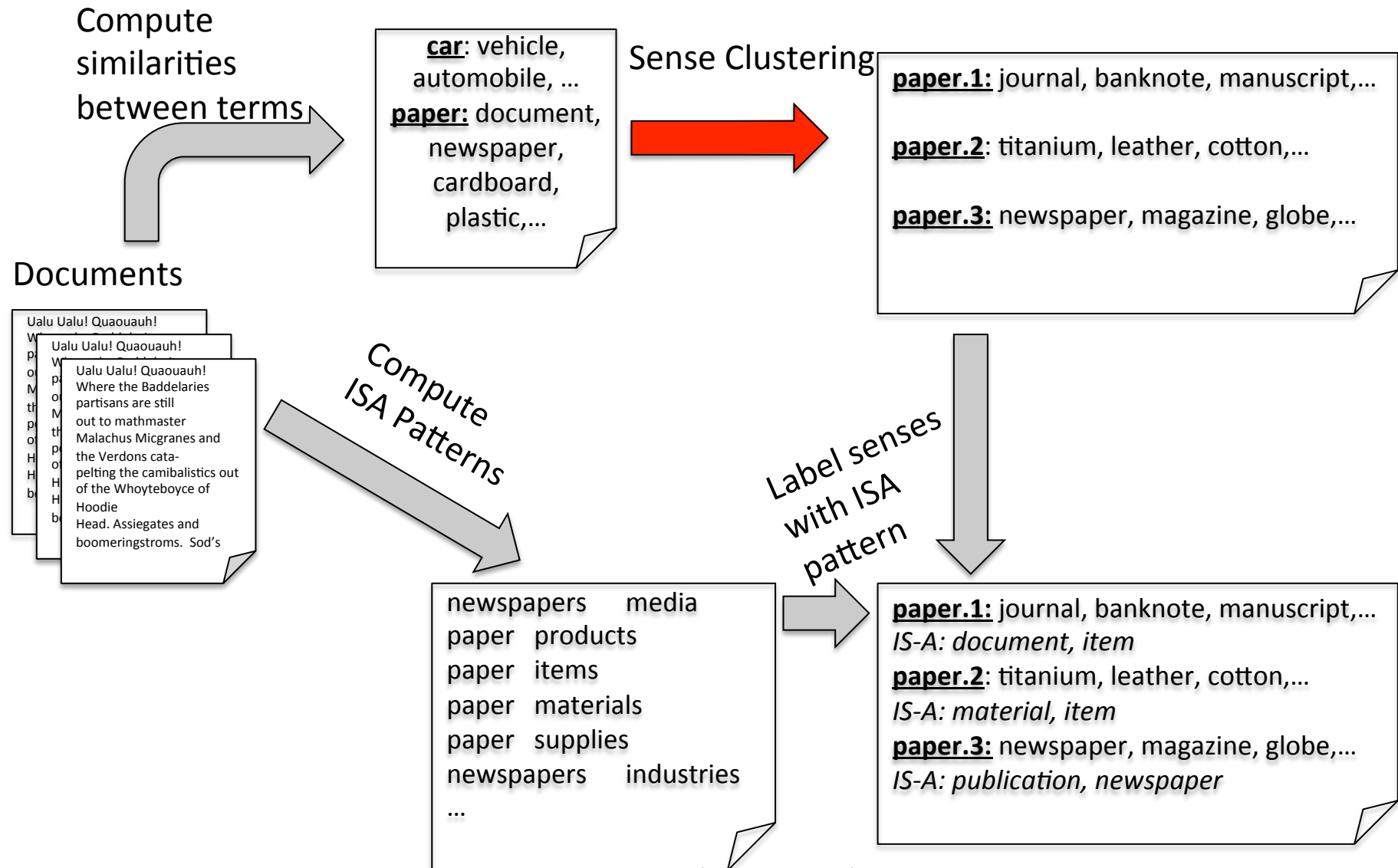
breast carcinoma	1000
breast cancer	186
carcinoma of the breast	84
human breast cancer	74
breast carcinomas	69
renal cell carcinoma	55
mammary carcinoma	55
breast tumor	54
malignant melanoma	52
breast cancers	46
melanoma	46
carcinomas	45

Comparing our method

- Evaluate against manually created thesaurus
- Compare:
 - Our method with two significance measures (LMI & PMI)
 - Two other approaches (Lin, Curran)
- Best Results with our method when using large corpora
- Our method scales better



The world of JoBimText

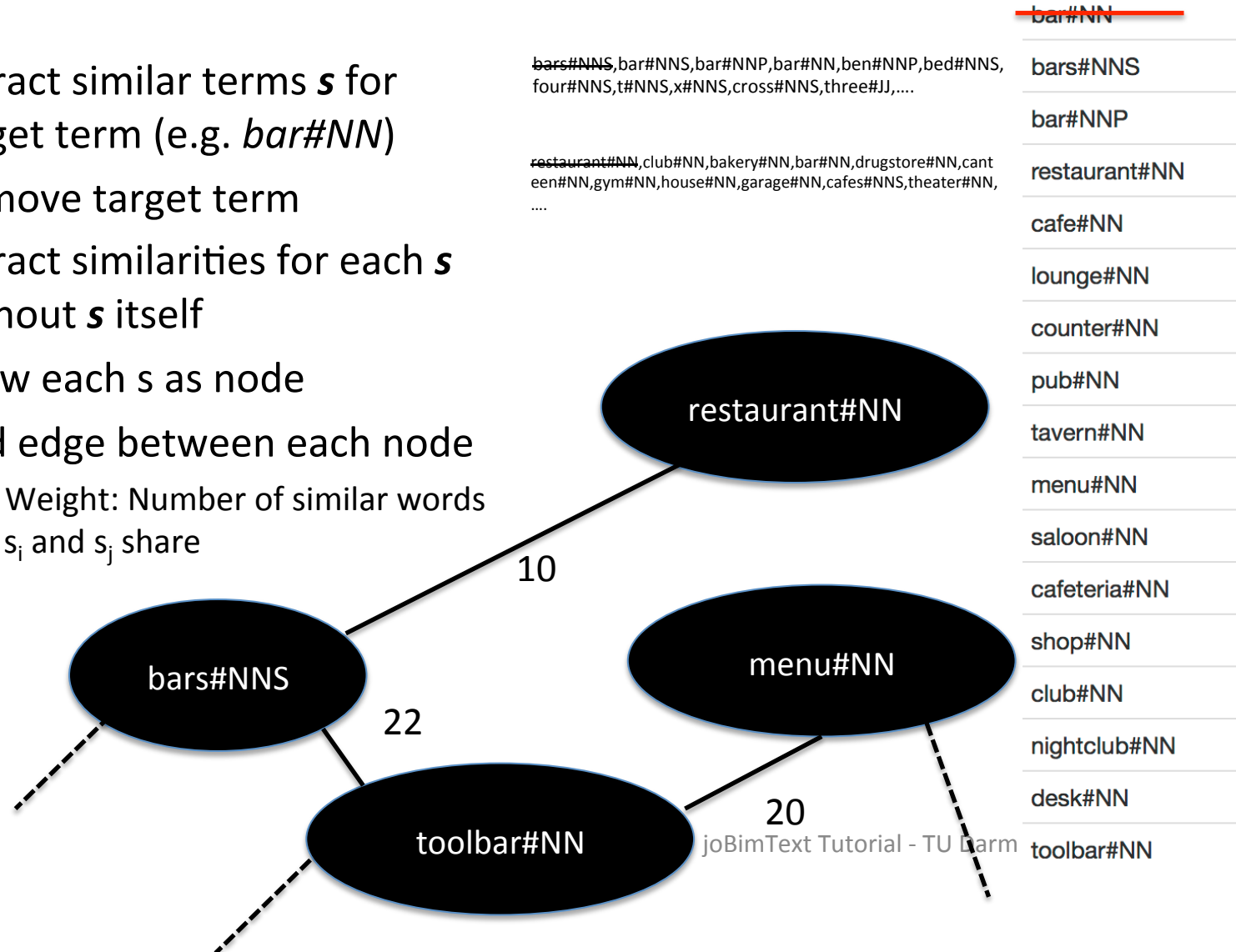


Computing Sense Cluster

- Use graph based Chinese Whispers
 - Pros:
 - No number of clusters needed
 - Time linear with the number of nodes
 - Contra:
 - non-deterministic
 - Does not converge

Building Graph

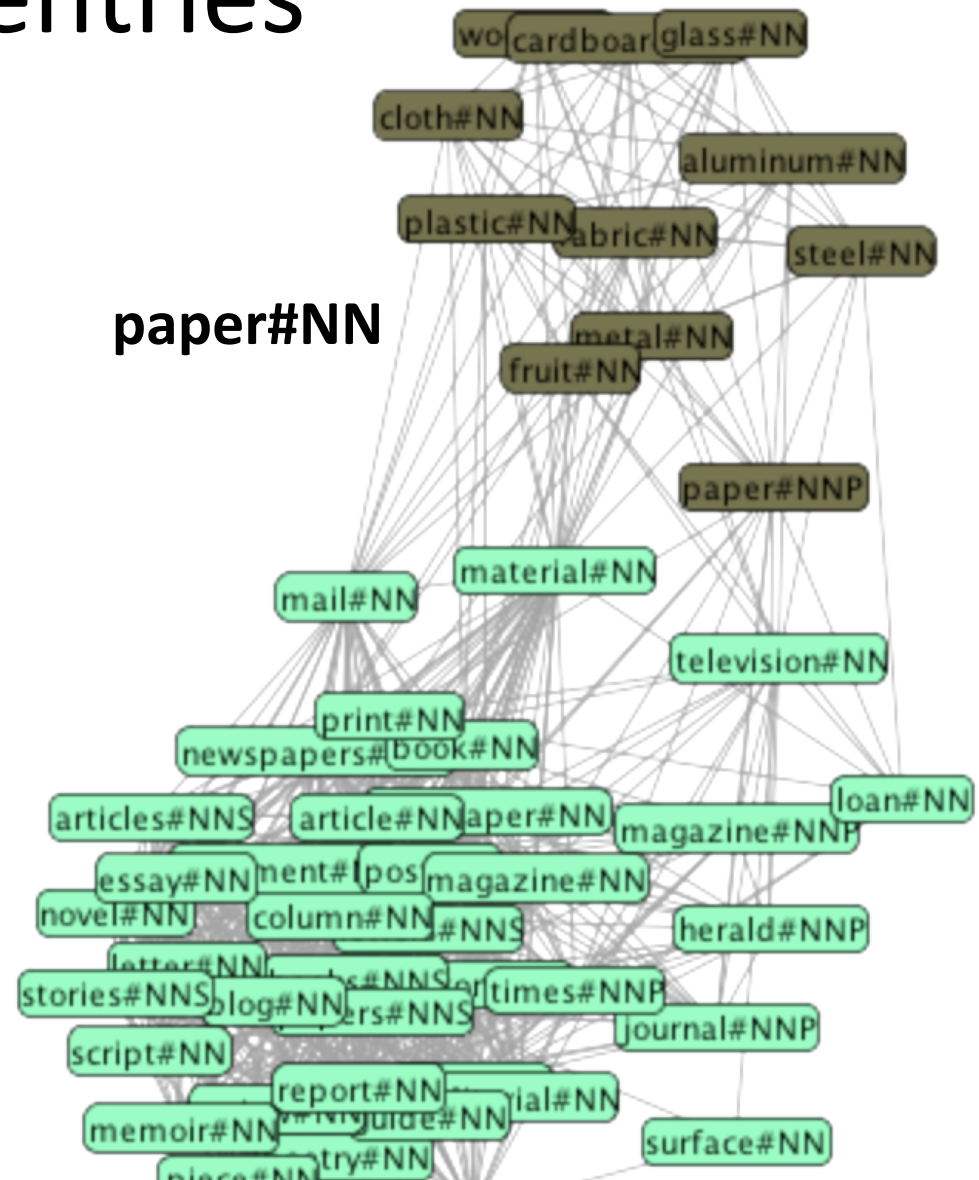
- Extract similar terms s for target term (e.g. *bar#NN*)
- Remove target term
- Extract similarities for each s without s itself
- Draw each s as node
- Add edge between each node
 - Weight: Number of similar words s_i and s_j share



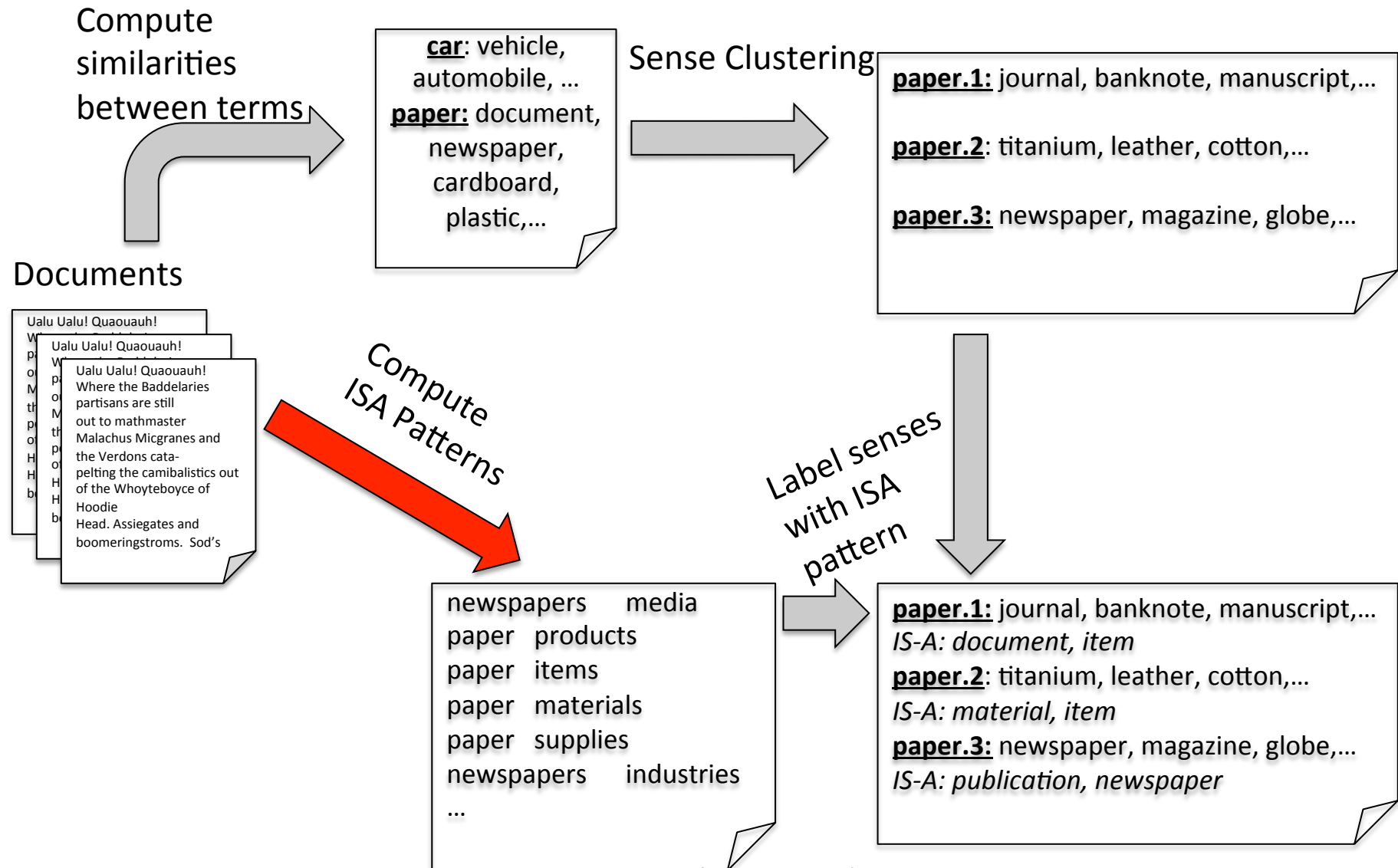
Clustering DT entries



paper#NN



The world of JoBimText



Cluster Labeling with IS-A Relations

- Run Hearst IS-A patterns (e.g. NP such as NP, NP and NP) on a large collection of text and store (noisy) IS-A pairs with their frequency, if above a threshold
- Activate hypernyms from cluster entries

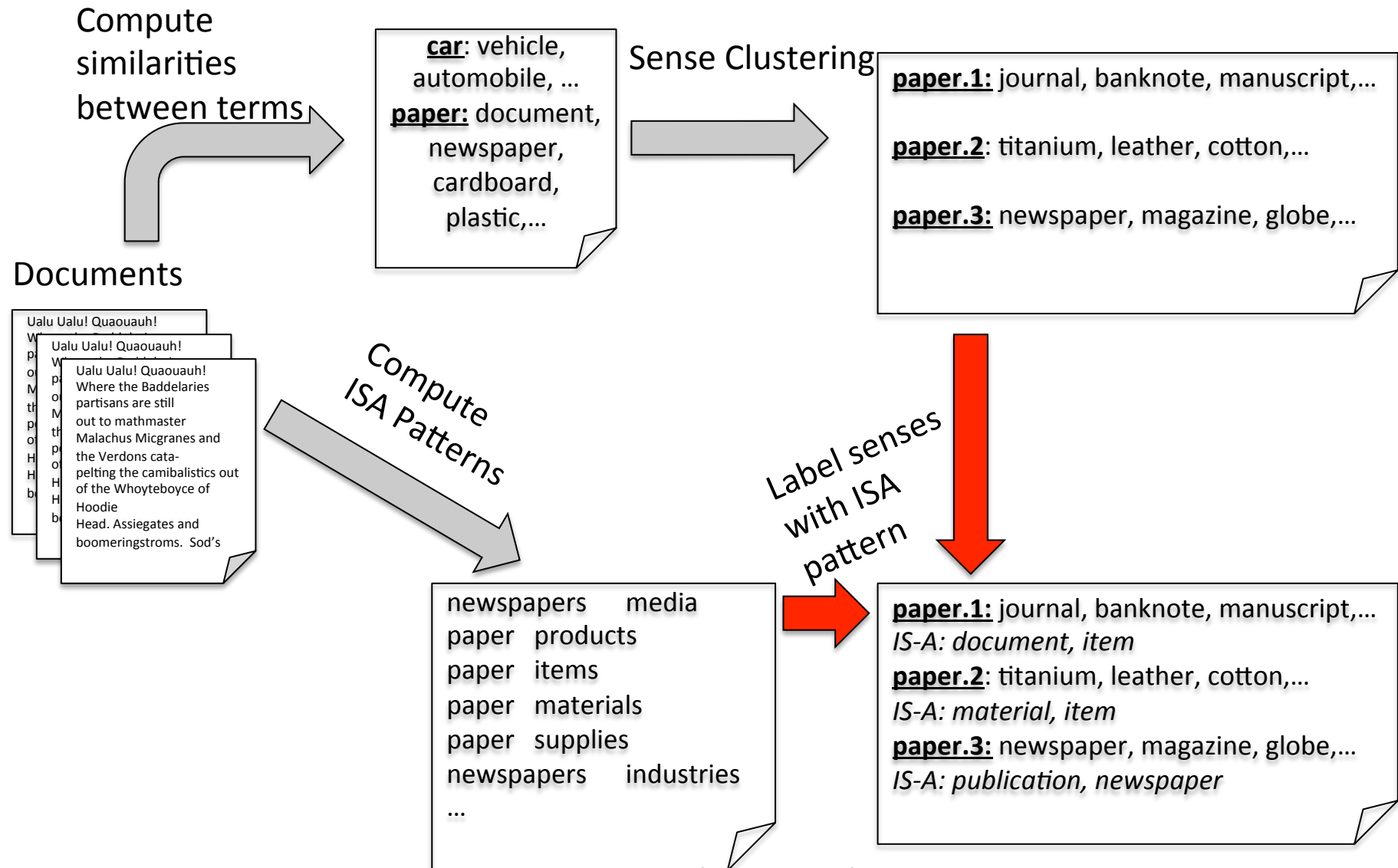
Typical regular polysemy in the medical domain:

influenza#0 viral gastroenteritis, bird flu, pulmonary anthrax, h1n1, tularaemia, west nile fever, mumps, influenza a, herpes zoster, respiratory infection, uri, phn, chicken pox, ...

influenza#1 trivalent influenza vaccine, antiviral, influenza vaccine, amantadine, peramivir, chemoprophylaxis, influenza vaccination, vaccine, poultry, chickenpox vaccine, ...

Hearst, M. Automatic Acquisition of Hyponyms from Large Text Corpora, Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, July 1992.

The world of JoBimText



Per-Cluster IS-A Pattern Counts

viral gastroenteritis ISA gastroenteritis 555	bird flu ISA flu 829	chicken pox ISA pox 2390
viral gastroenteritis ISA illness 27	bird flu ISA avian influenza 42	chicken pox ISA virus 93
viral gastroenteritis ISA stomach flu 24	bird flu ISA influenza 27	chicken pox ISA infection 76
viral gastroenteritis ISA cause 15	bird flu ISA avian flu 16	chicken pox ISA symptom 43
viral gastroenteritis ISA T		chicken pox ISA illness 38

**infection(3310937) disease(1748000) virus(817950) cause(783692)
 symptom(578480) fever(375228) condition(209022) illness(192675)
 complication(161158) influenza(155469)**

influenza#0 viral gastroenteritis, bird flu, pulmonary anthrax, h1n1, tularaemia, west nile fever, mumps, influenza a, herpes zoster, respiratory infection, uri, phn, chicken pox, ..

influenza#1 trivalent influenza vaccine, antiviral, influenza vaccine, amantadine, peramivir, chemoprophylaxis, influenza vaccination, vaccine, poultry, chickenpox vaccine, ...

peramivir ISA inhibitor 4	vaccine(38566) drug(9990) agent(5004) vaccination(3960) treatment(2796)	influenza vaccine ISA vaccine 1532
peramivir ISA option 4	inhibitor(1504) medication(792) oseltamivir(752) medicine(448)	influenza vaccine ISA antiviral 10
peramivir ISA neuraminidase inhibitor 4	zanamivir(396)	influenza vaccine ISA intranasal 8
		influenza vaccine ISA 2010-XX-XX 8
		influenza vaccine ISA LAIV 8
		influenza vaccine ISA table 8
		influenza vaccine ISA contrast 6

- Sum counts of ISA hypernym per cluster
- Multiply by number of times it was found by the cluster members

Gliozzo A., Biemann C., Riedl M., Coppola B., Glass M. R., Hatem M. (2013): JoBimText Visualizer: A Graph-based Approach to Contextualizing Distributional Similarity. Proceedings of the 8th Workshop on TextGraphs in conjunction with EMNLP 2013, Seattle, WA, USA

What's next

- Word Sense Induction System
 - Align sense cluster to word within sentence
- Semantic Search Engine
- Collapsed dependencies (for English and German)
- Multiword Expression Detector

NF-kappa B
transcription factors
transcription factor
I kappa B alpha
activated T cells
nuclear factor
human monocytes
gene expression
T lymphocytes

hausse des prix
mise en oeuvre
prise de participation
chiffre d'affaires
formation professionnelle
population active
taux d'intérêt
politique monétaire
Etats-Unis

APPLICATIONS FOR JOBIMTEXT

Solving OOV problem for POS tagging

- POS-tagging is hard for unknown words, which have not been observed in the training
- Schema: replace the unknown words with the most similar known word from an n-gram DT
- tag new sentence with standard POS tagger

Renting out an JJ unfurnished JJ one-bedroom NN triplex in San Francisco
empty two-bedroom duplex
three-bedroom
two-room

OOV acc. PTB, TreeTagger
1-dimensional: 37.8%
2-dimensional: 74.1%

Knowledge-based Word Sense Disambiguation (Lesk-style)

A patient **fell** over a **stack** of magazines in an aisle at a physiotherapist practice.

customer	rose	pile		field	physician	session
student	dropped	copy		hill	attorney	game
individual	climbed	lots		line	psychiatrist	camp
person	increased	dozens		river	scholar	workouts
mother	slipped	array		stairs	engineer	training
user	declined	collection		road	journalist	meeting
passenger	tumbled	amount		hall	contractor	work
..	surged	ton		driveway	...	...
...	...	...		...	...	...

Zero word overlap

WordNet: S: (n) magazine (product consisting of a paperback periodic publication as a physical object) "tripped over a **pile** of magazines"

jumped	stack
woke	tons
turned	piece
drove	heap
walked	collection
blew	bag
put	loads
fell	mountain

Overlap = 2

Overlap = 1

Overlap = 2

Tristan Miller, Chris Biemann, Torsten Zesch, Iryna Gureych (2012): Using Distributional Similarity for Lexical Expansion in Knowledge-based Word Sense Disambiguation. Proceedings of COLING-12, Mumbai, India

Lexical Substitution using JobimText

Given: Sentences

This book is more than just a compendium of conference papers, however

Target Word



Goal:

Find substitutions for target word that fit into the context

Evaluation:

Compare substitutions against gold standard
(several measures exist)

document 2;presentation 1;treatise 1;article 1;manuscript 1;



#Annotators

Lexical Substitution Results

Ablation test for two different domain datasets

open domain [1]

	GAP	P@1
w/o n-gram features	47.3	48.9
w/o distr. thesaurus	49.8	55.0
w/o POS features	51.6	56.3
w/o WN features	51.7	57.0
Our model (all)	52.4	57.7

medical domain [2]

System	MAP	P@1
Baseline	0.6408	0.5365
ALL	0.7048	0.6366
w/o DT	0.5798	0.4835
w/o UMLS	0.6618	0.5651
w/o Ngrams	0.7009	0.6252
w/o POS	0.7027	0.6323

Ranking
using only
DT

- significantly improvement with DT

[1] Supervised All-Words Lexical Substitution using Delexicalized Features György Szarvas and Chris Biemann and Iryna Gurevych In Proceedings of NAACL-HLT 2013, Seattle, USA

[2] Martin Riedl, Michael R. Glass, Alfio Gliozzo, 2014, Lexical Substitution for the Medical Domain, In Proceedings of EMNLP 2014, Doha, Qatar

Any Questions or Comments?

question#NN

query#NN

doubt#NN

concern#NN

issue#NN

complaint#NN

dilemma#NN

idea#NN

uncertainty#NN

matter#NN

comment#NN

remark#NN

suggestion#NN

statement#NN

commentary#NN

feedback#NN

assertion#NN

announcement#NN

speech#NN

criticism#NN

Timetable

- 11:00 – 12:00 Methods and Applications of JoBimText 
- 12:00 – 13:00 Using JobimText I
- 13:00 – 14:00 Lunch
- 14:00 – 14:30 Using JoBimText II
- [14:30 – 15:00 Discussion]