



TECHNISCHE
UNIVERSITÄT
DARMSTADT

JoBimText: A framework for distributional semantics

6/11/2015

By

Martin Riedl and Eugen Ruppert



LT@TU Darmstadt

<http://www.lt.informatik.tu-darmstadt.de/>

People

Prof. Dr. Chris Biemann



Dr. Alexander Panchenko



Dr. Bonaventura Coppola



Sarah N. Kohail



Benjamin Milde



Stephan Radeck-Arneth



Steffen Remus



Martin Riedl



Eugen Ruppert

Gerold Hintz

Petra Stegmann

Seid Muhie Yimam

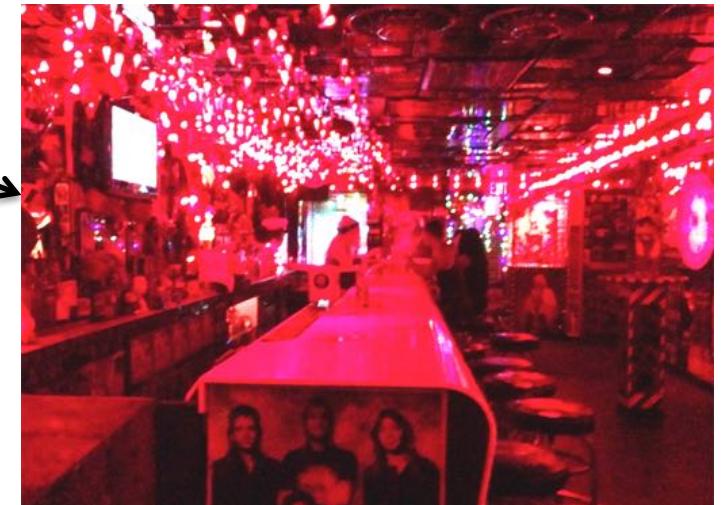


Timetable

- 14:00 – 14:30 Methods and Applications of JoBimText
- 14:30 – 16:00 Access Semantic Models
- 16:15 – 17:30 Computing Semantic Models
- 17:30 – Discussion @ the Beer garden

Understanding Text

The bar serves delicious beer

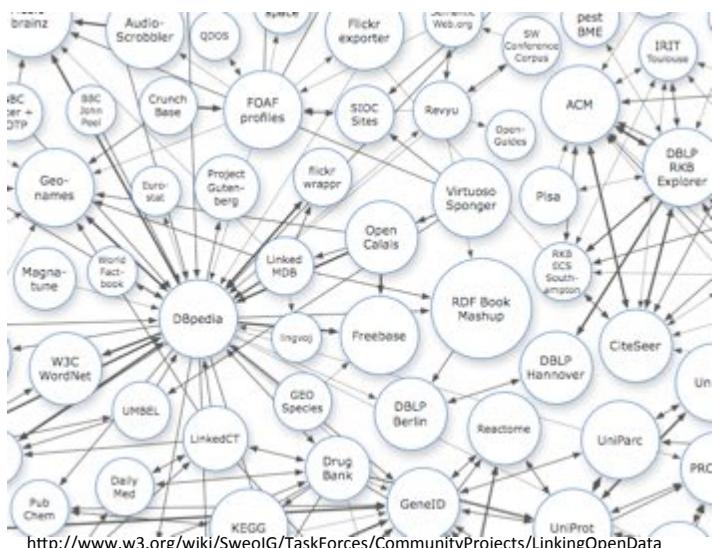


Click on Mail in the menu bar

Use Dictionaries or Ontologies



Photo by zeh fernando under Creative Commons licence



<http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

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...”*

Motivation for JoBimText



“Structure from nothing,
get your knowledge for free”



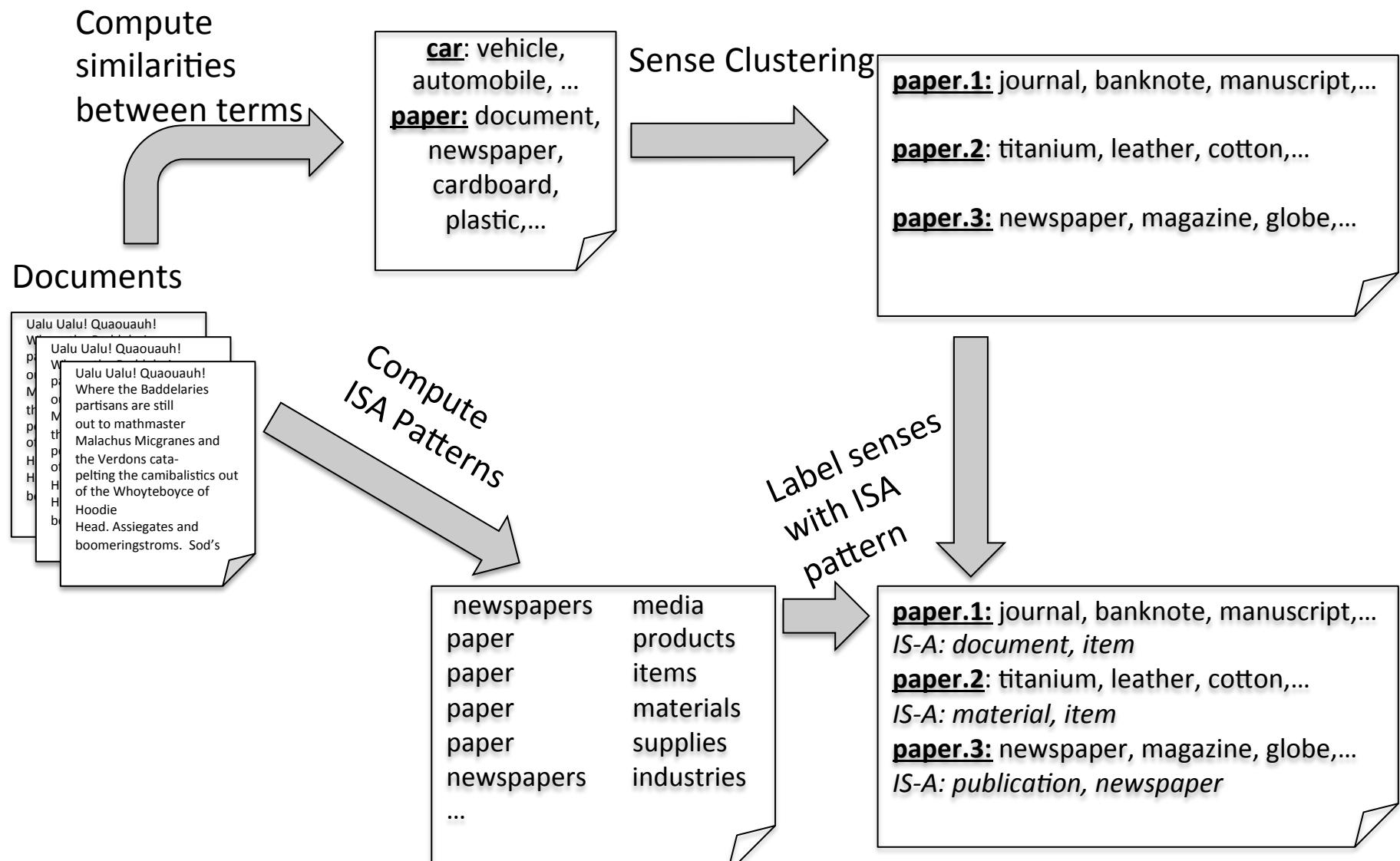
- Combine unsupervised distributional method
- Use Methods which scale to large data
- Provide everything for free (source, models)

Distributional Semantics

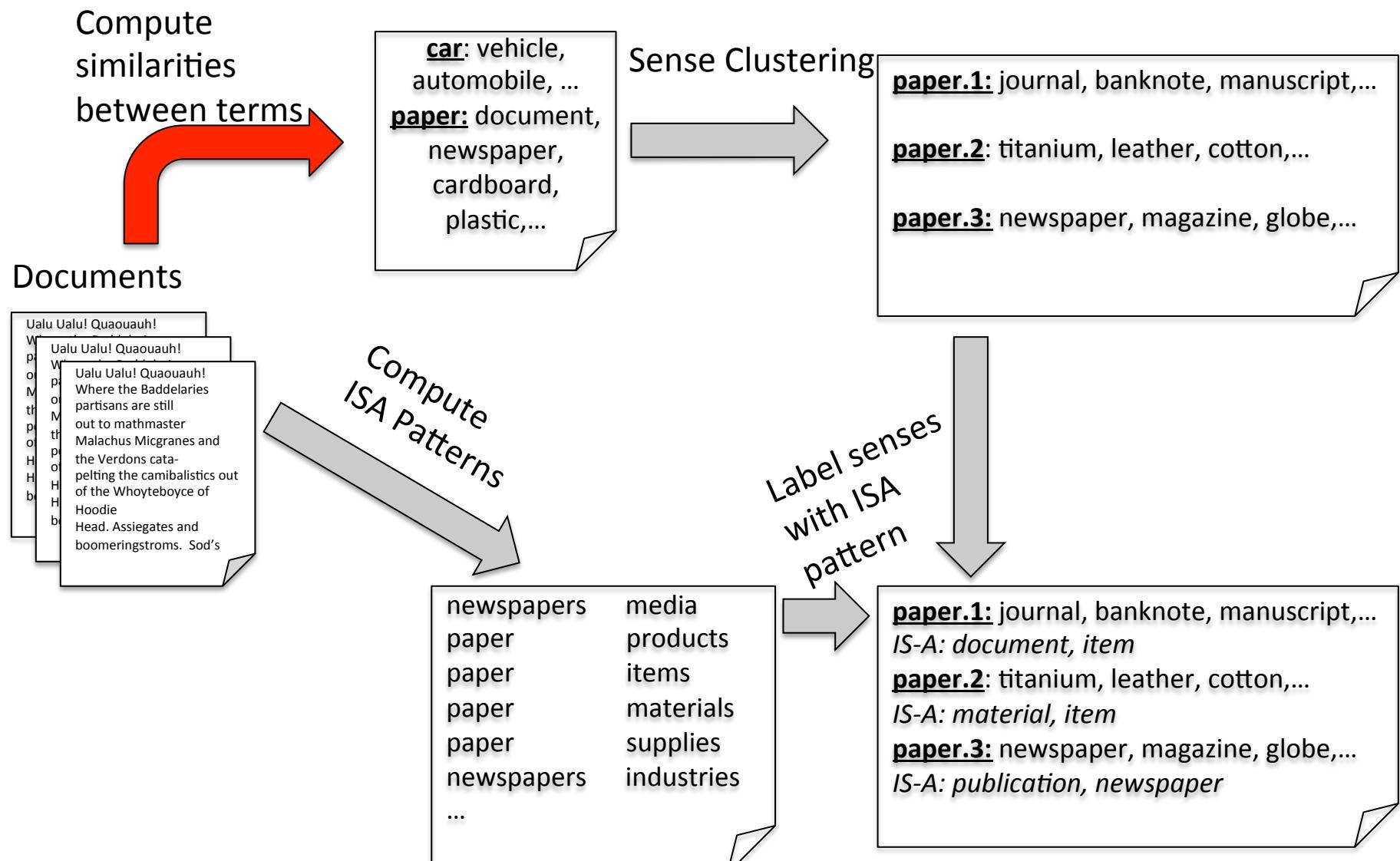
- Based on the distributional hypothesis
(popularized by First 1957)

“a word is characterized
by the company it keeps”

The world of JoBimText



The world of JoBimText



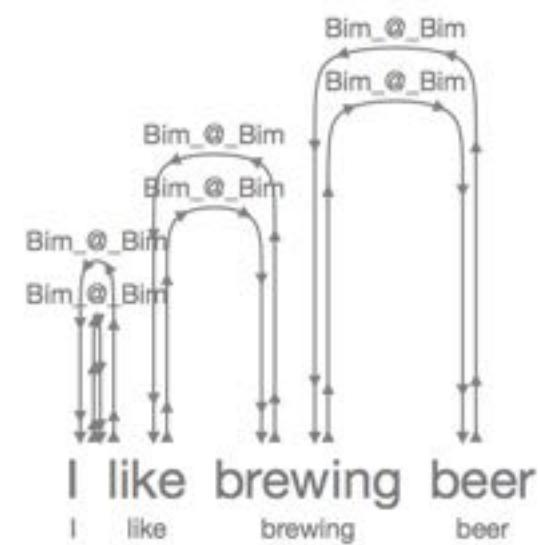
Computing Similarities

- @@-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

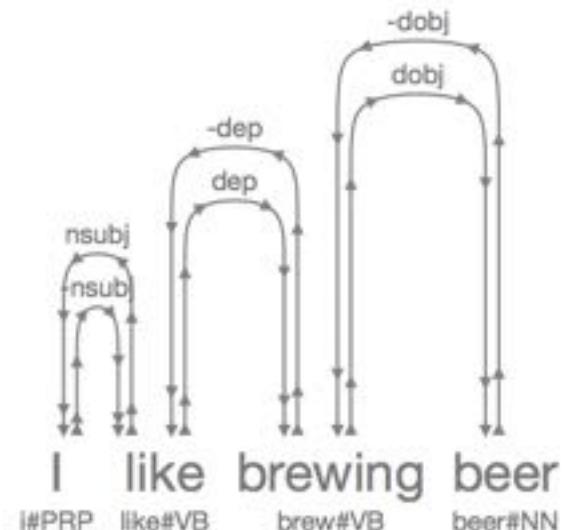
- Input: I like brewing beer.
- Holing Operation:
 - Extract Jo – Bim

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



Example: Parsing @@-Operation

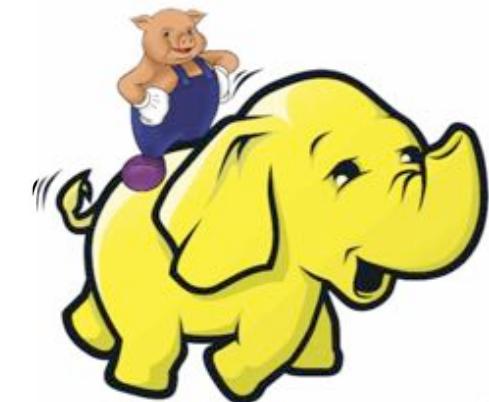
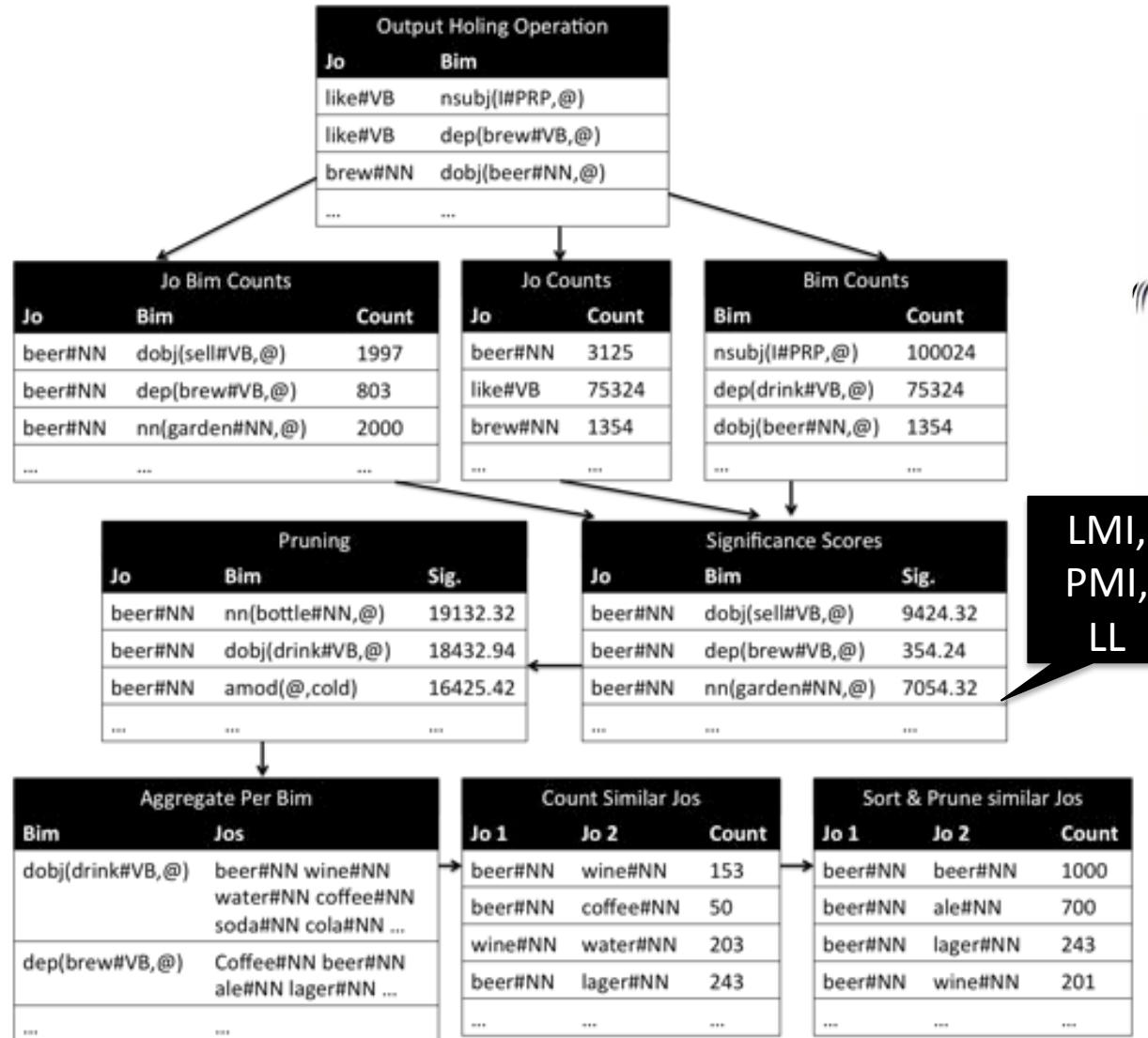
- Input: I like brewing beer.
- Holing Operation:
 - Parsing and Lemmatization
 - Extract Jo – Bim



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



LMI,
PMI,
LL

Hadoop
Pig

Distributional Thesaurus based on Jo's

Similar terms for
beer using
dependency
parses on
Newspaper data

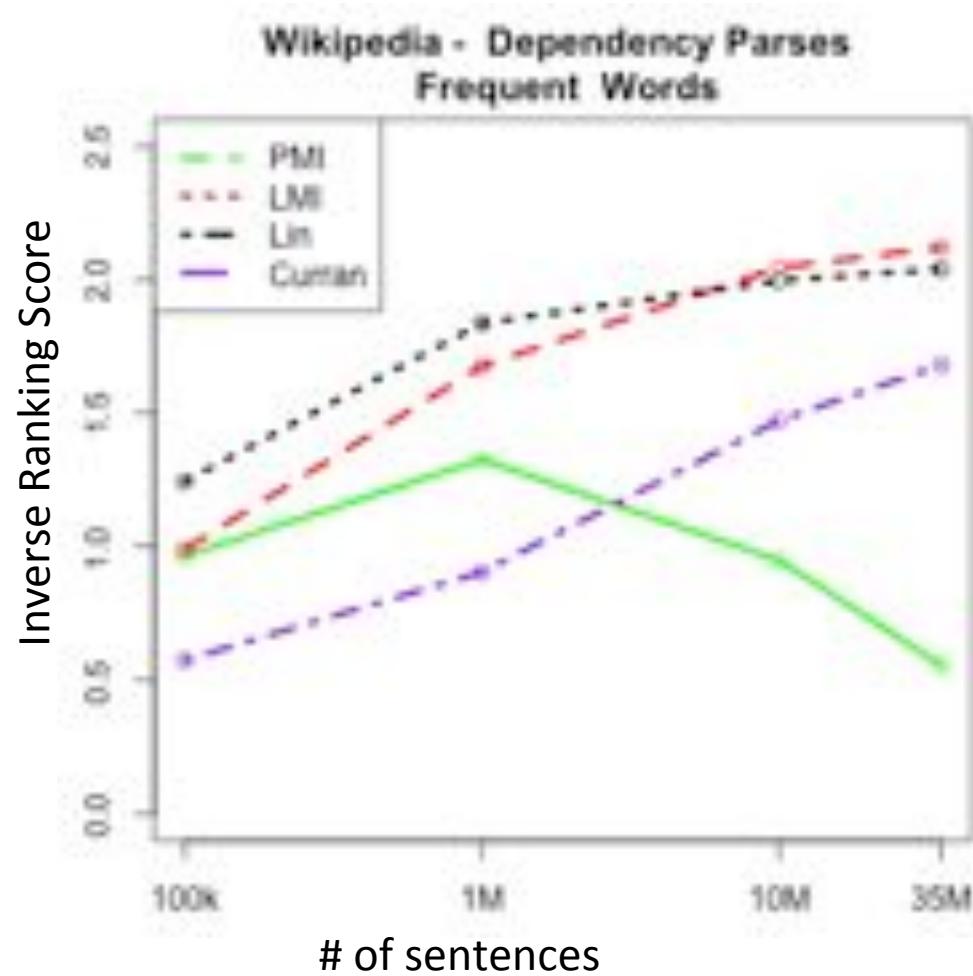
beer#NN	711
drink#NN	190
wine#NN	183
soda#NN	179
coffee#NN	175
beverage#NN	168
liquor#NN	155
tea#NN	154
lager#NN	140
champagne#NN	138
vodka#NN	136
whiskey#NN	126
ale#NN	126

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

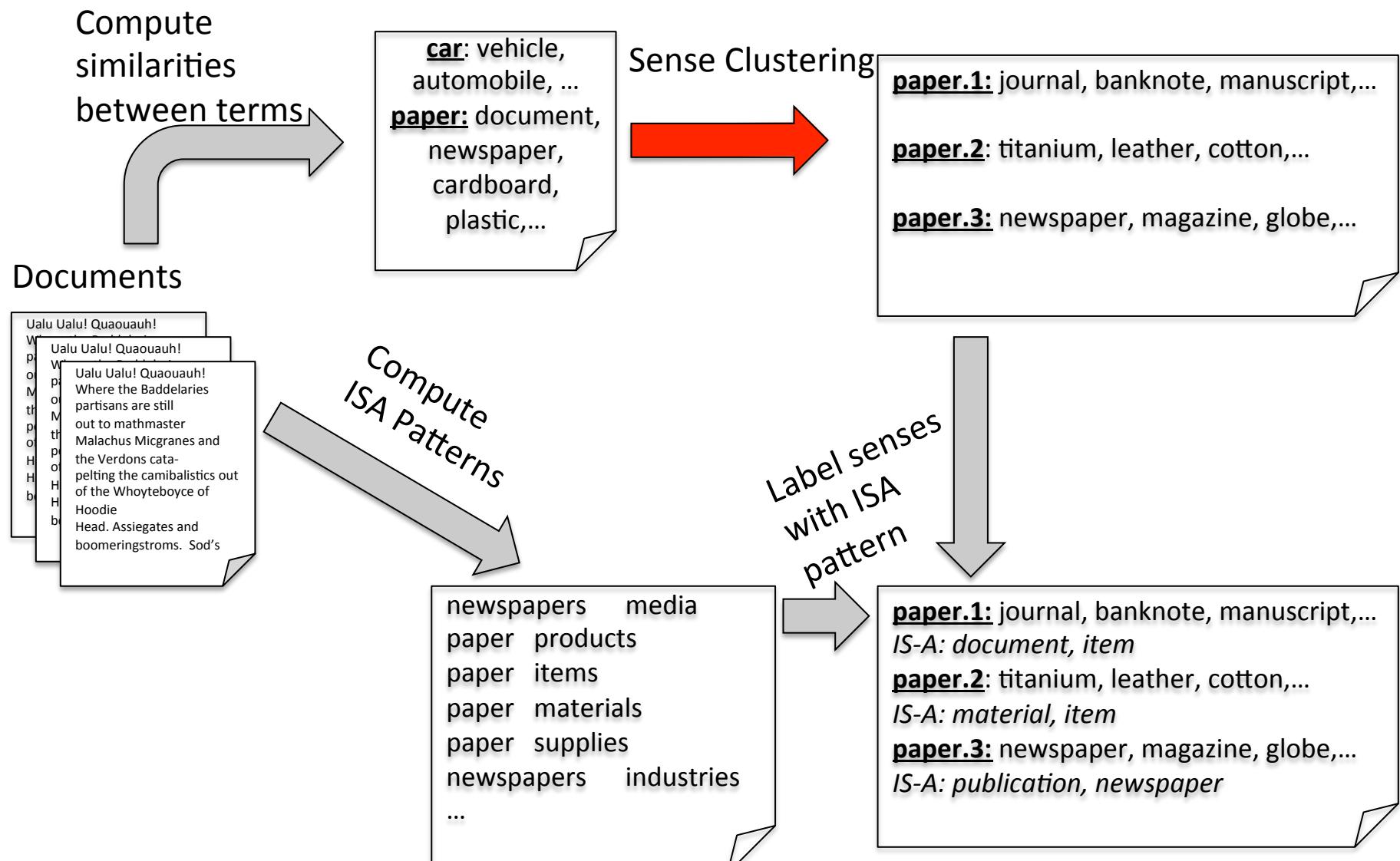
red blood cells	1000
erythrocytes	167
red cells	101
RBCs	43
human erythrocytes	33
and goats	27
platelets	27
peripheral blood lymphocytes	25
and cattle	24
RBC	23
red blood cell	22
reticulocytes	22
faeces	20
cell volume	20

Performance of 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



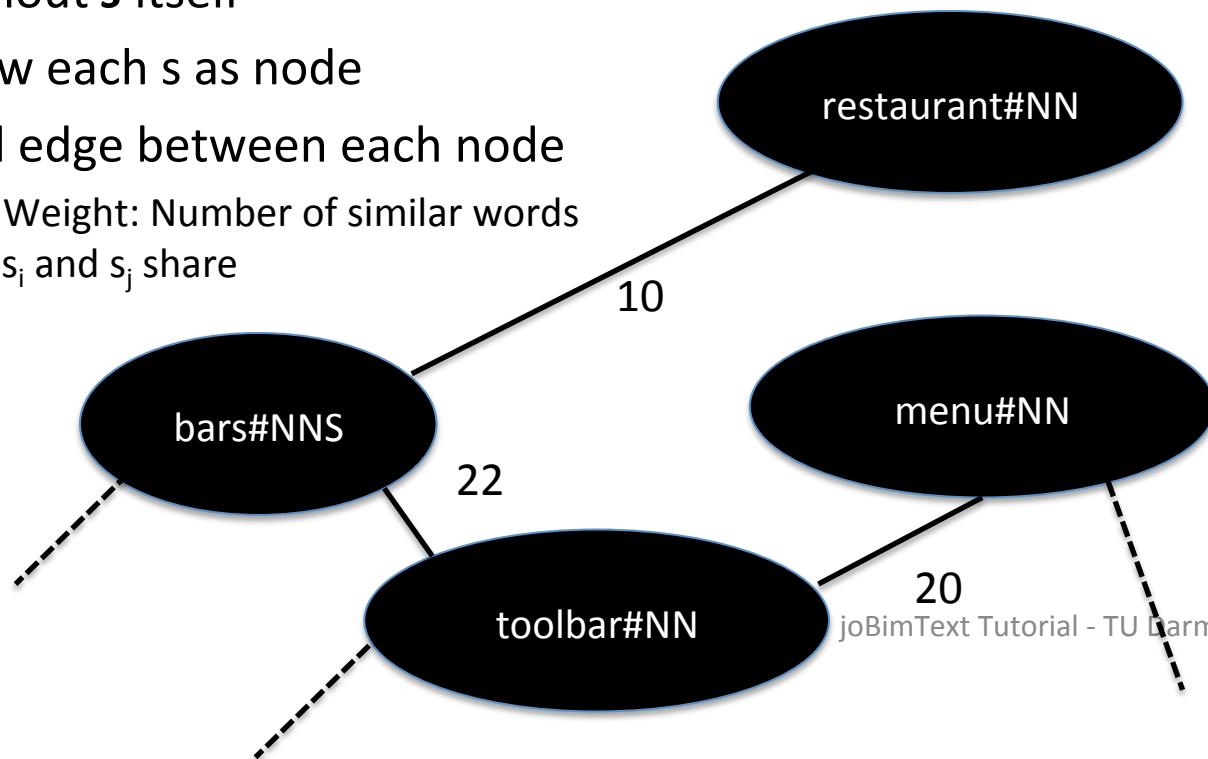
Computing Sense Cluster

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

C. Biemann (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.

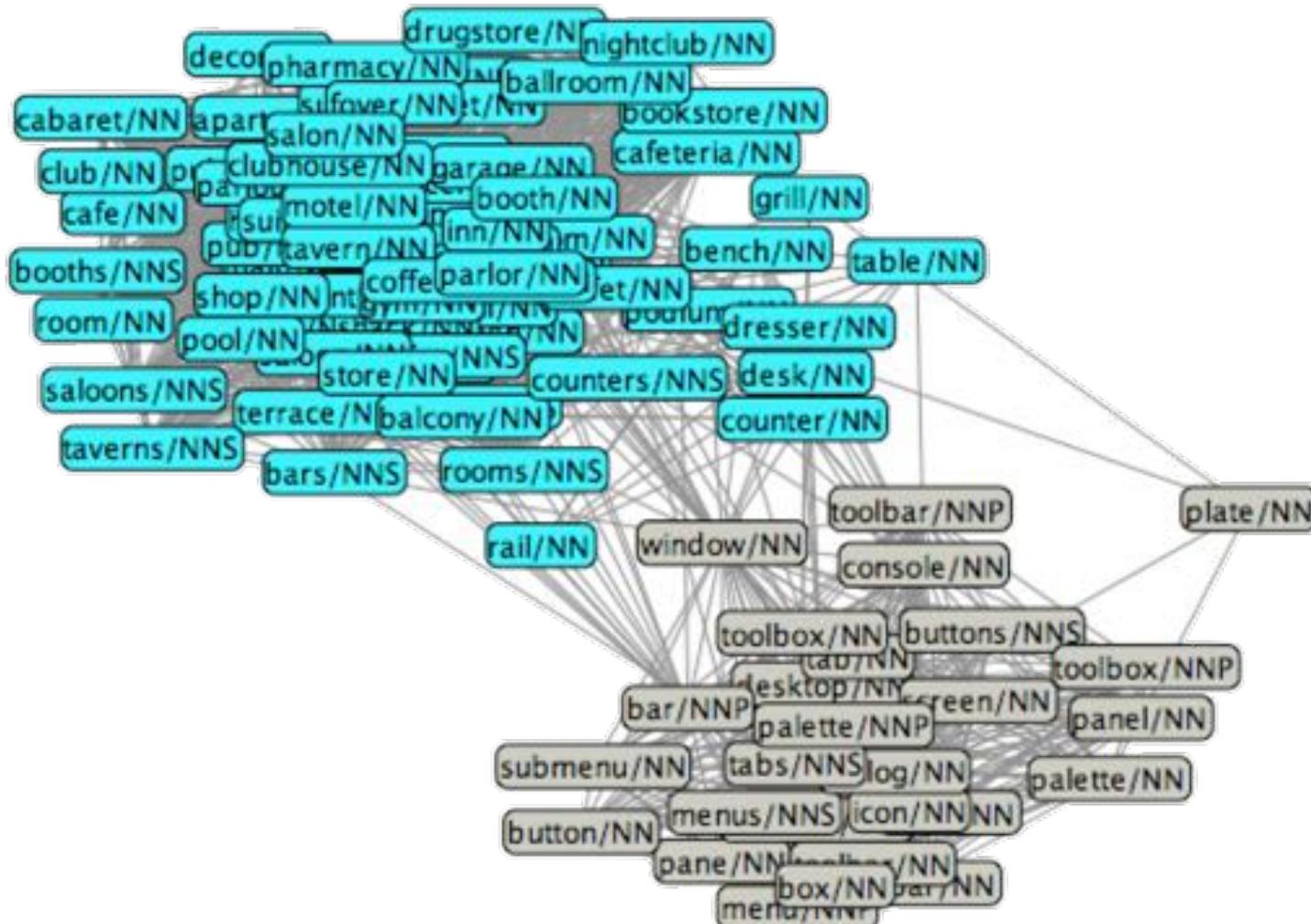
Building Graph

- 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



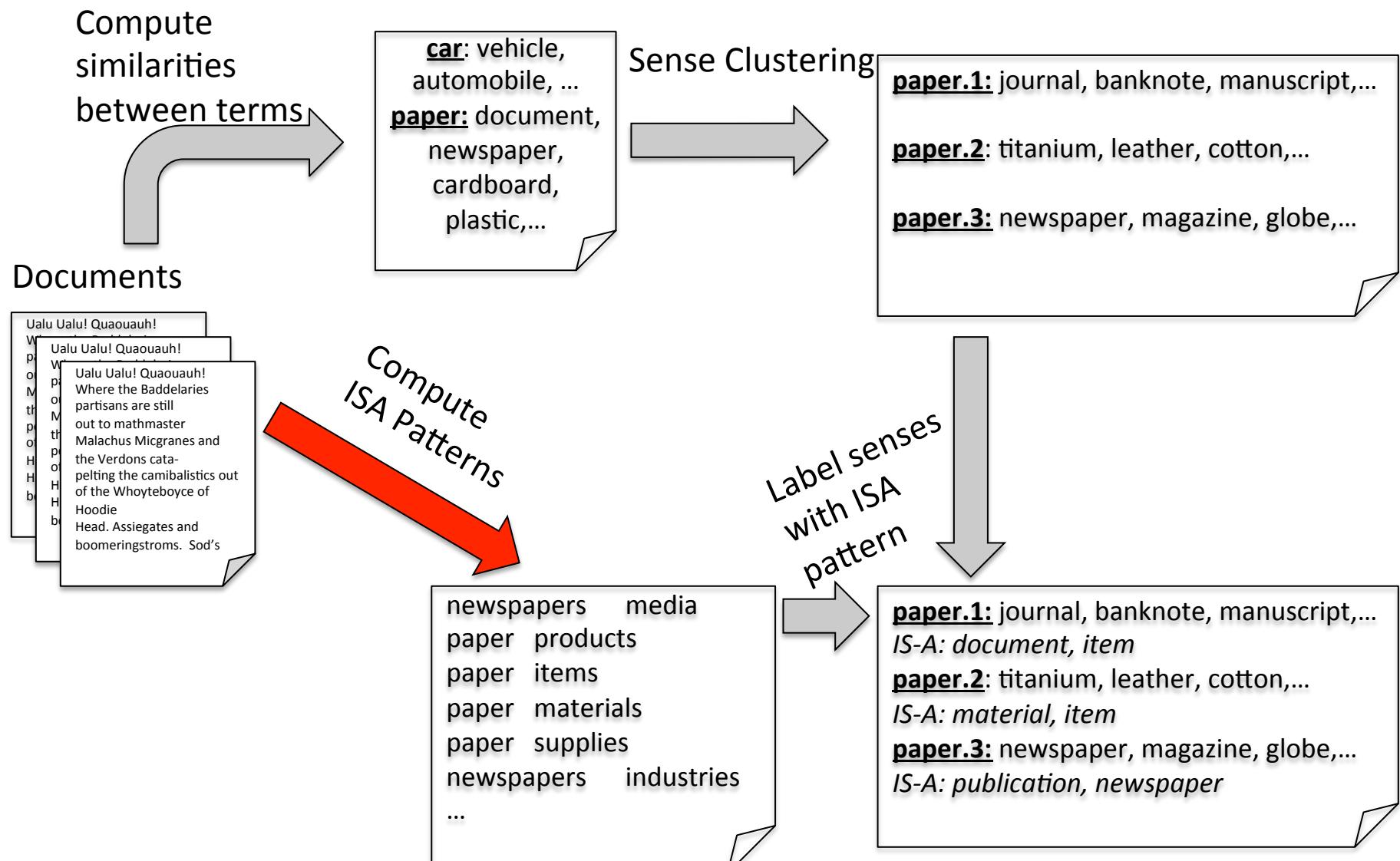
<i>bar#NN</i>
bars#NNS
bar#NNP
restaurant#NN
cafe#NN
lounge#NN
counter#NN
pub#NN
tavern#NN
menu#NN
saloon#NN
cafeteria#NN
shop#NN
club#NN
nightclub#NN
desk#NN
toolbar#NN

Clustering DT entries



bar#NN

The world of JoBimText



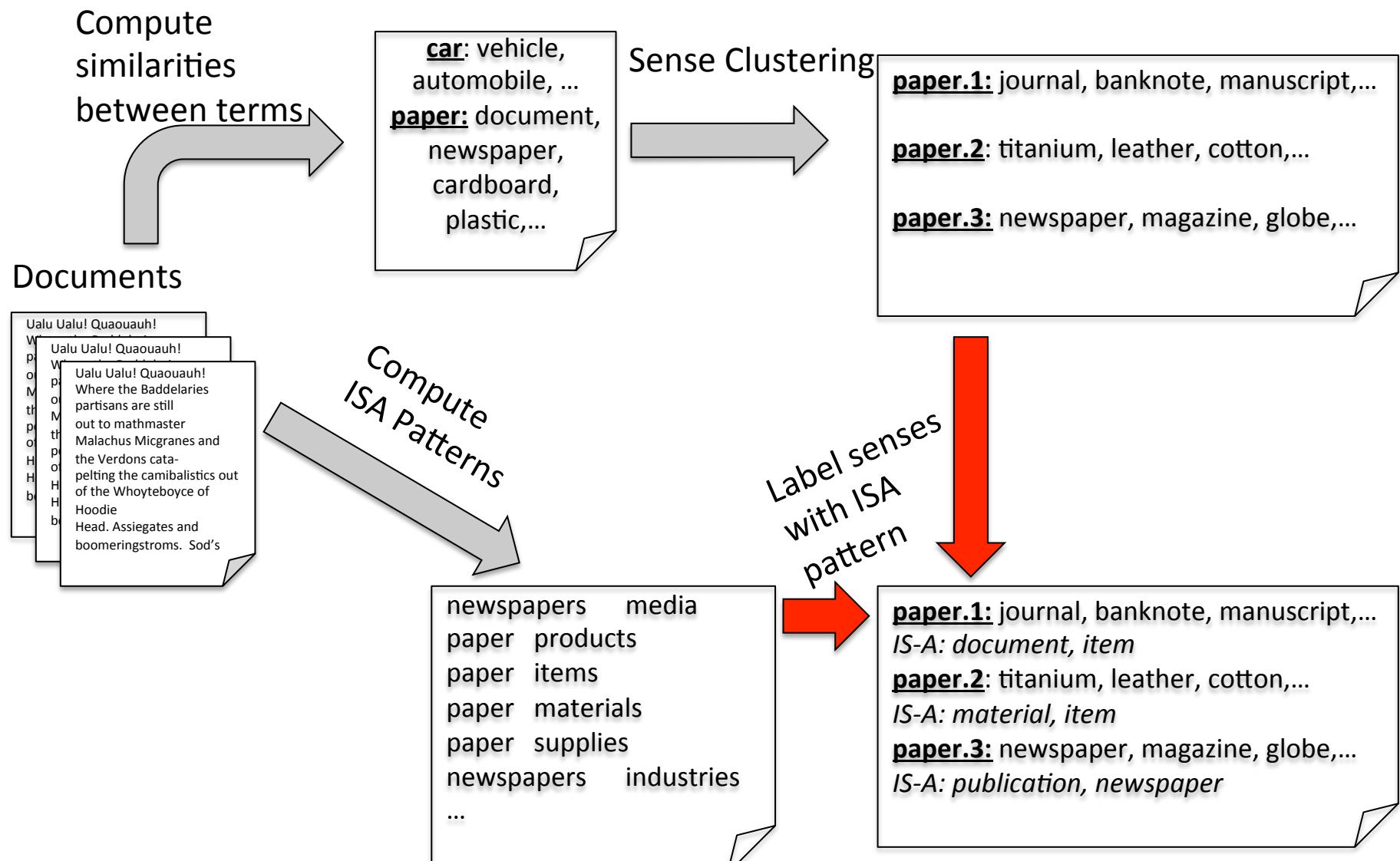
Cluster Labeling with IS-A Relations

- 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

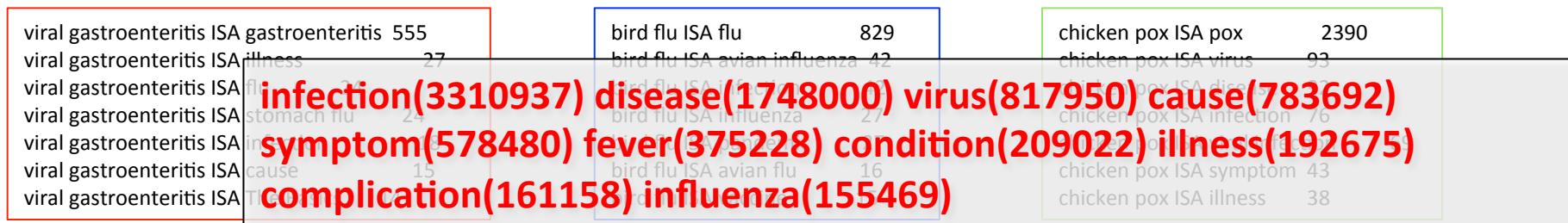
bird flu ISA flu	829
bird flu ISA avian influenza	42
bird flu ISA infection	42
bird flu ISA influenza	27
bird flu ISA pandemic	27
bird flu ISA avian flu	16
bird flu ISA vaccine	15

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



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, ...



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

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 **unfurnished** **one-bedroom** **triplex** in San Francisco
empty two-bedroom duplex
three-bedroom
two-room

JJ

JJ

NN

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
student	dropped	copy
individual	climbed	lots
person	increased	dozens
mother	slipped	array
user	declined	collection
passenger	tumbled	amount
..	surged	ton
...

Zero word overlap

field	physician	session
hill	attorney	game
line	psychiatrist	camp
river	scholar	workouts
stairs	engineer	training
road	journalist	meeting
hall	contractor	work
driveway
...

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 Gurevych (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	comment#NN
query#NN	remark#NN
doubt#NN	suggestion#NN
concern#NN	statement#NN
issue#NN	commentary#NN
complaint#NN	feedback#NN
dilemma#NN	assertion#NN
idea#NN	announcement#NN
uncertainty#NN	speech#NN
matter#NN	criticism#NN