JoBimText Framework for Distributional Semantics

Alexander Panchenko
TU Darmstadt — FG Language Technology

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Most slides by

Martin Riedl

&

Eugen Ruppert

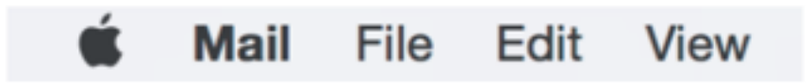
from TU Darmstadt

Plan

- **Distributional Similarity**
- **Word Sense Induction**
- **Word Sense Disambiguation**

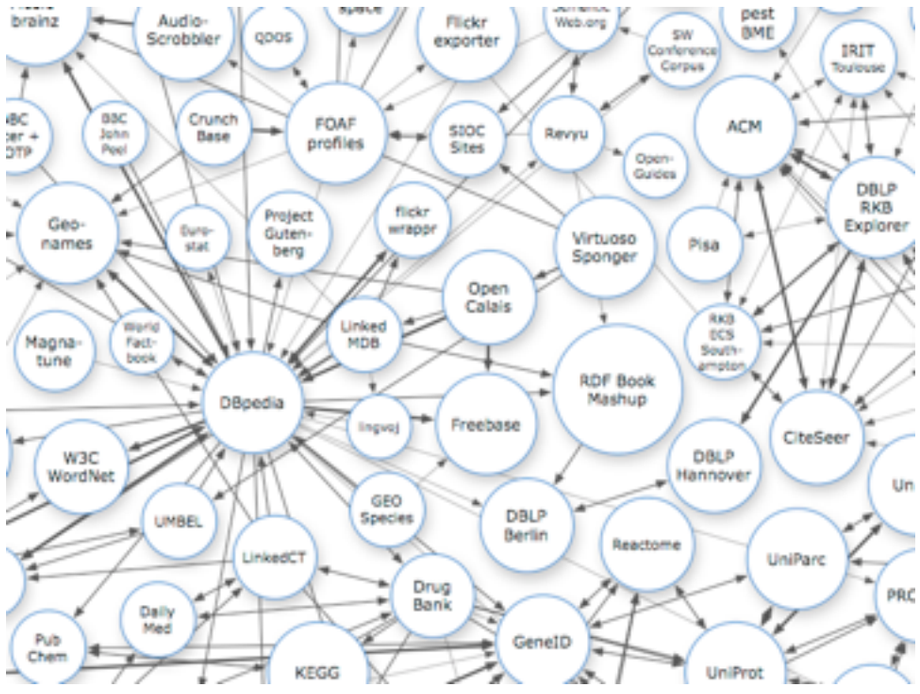
Motivation: Text Understanding

The bar serves delicious beer



Click on Mail in the menu bar

Why Not To Use Dictionaries or Ontologies



Advantages:

- Sense inventory given
- Linking to concepts
- Full control

Disadvantages:

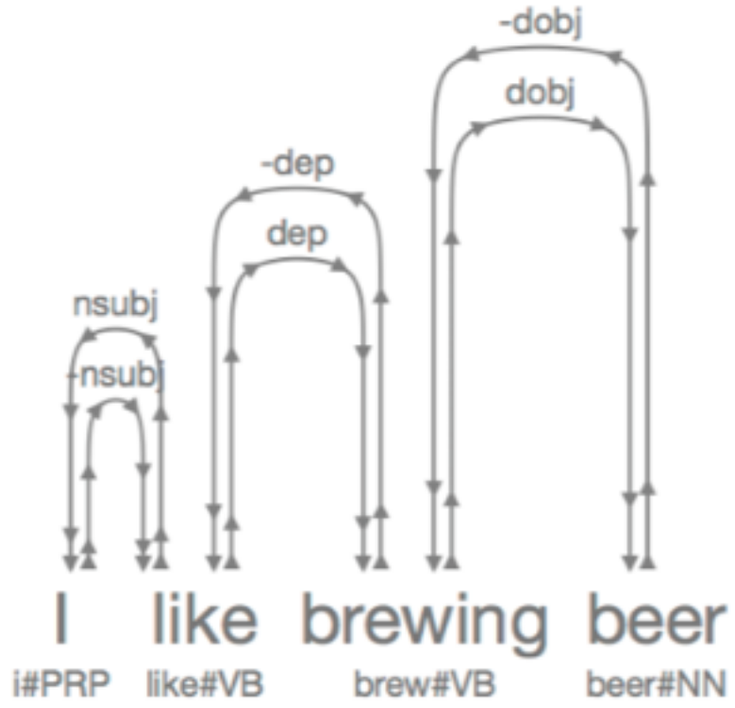
- Dictionaries have to be created
- Dictionaries are incomplete
- Language changes constantly: new words, new meanings …

"give a man a fish and you feed him for a day…"

Distributional Similarity

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)

ID	Sentence
1	Nothing is gained.
2	Similiarity of words ...
...	...

Context Feature Extractor

Language Element	Context Feature	Count
hard#a	(adj_mod; @; cheese#n)	34
cheese#n	(adj_mod; Gouda-like#a;@)	12
...	...	...

Language Element Count

Language Element	Count
cheese#n	70
hard#n	40
...	...

Context Feature Count

Context Feature	Count
(adj_mod; @; cheese#n)	50
(adj_mod; hard#a;@)	30
...	...

Frequency Significance Measure

Language Element	Context Feature	Sign.
hard#a	(adj_mod; @; cheese#n)	15.7
cheese#n	(adj_mod; Gouda-like#a; @)	7.3
...	...	...

Pruning

Language Element	Context Feature	Sign.
hard#a	(adj_mod;@;cheese#n)	15.7
cheese#n	(adj_mod; yellow#a#; @)	17.3
...	...	...

Aggregate Per Feature

Context Feature	Language Elements
(adj_mod;@;cheese#n)	hard#a; yellow#a; french#a
(adj_mod; hard#a#; @)	cheese#n; stone#n
...	...

Similarity Count

Language Element 1	Language Element 2	Score
hard#a	yellow#a	50
cheese#n	stone#n	90
...	...	...

Similarity Sort

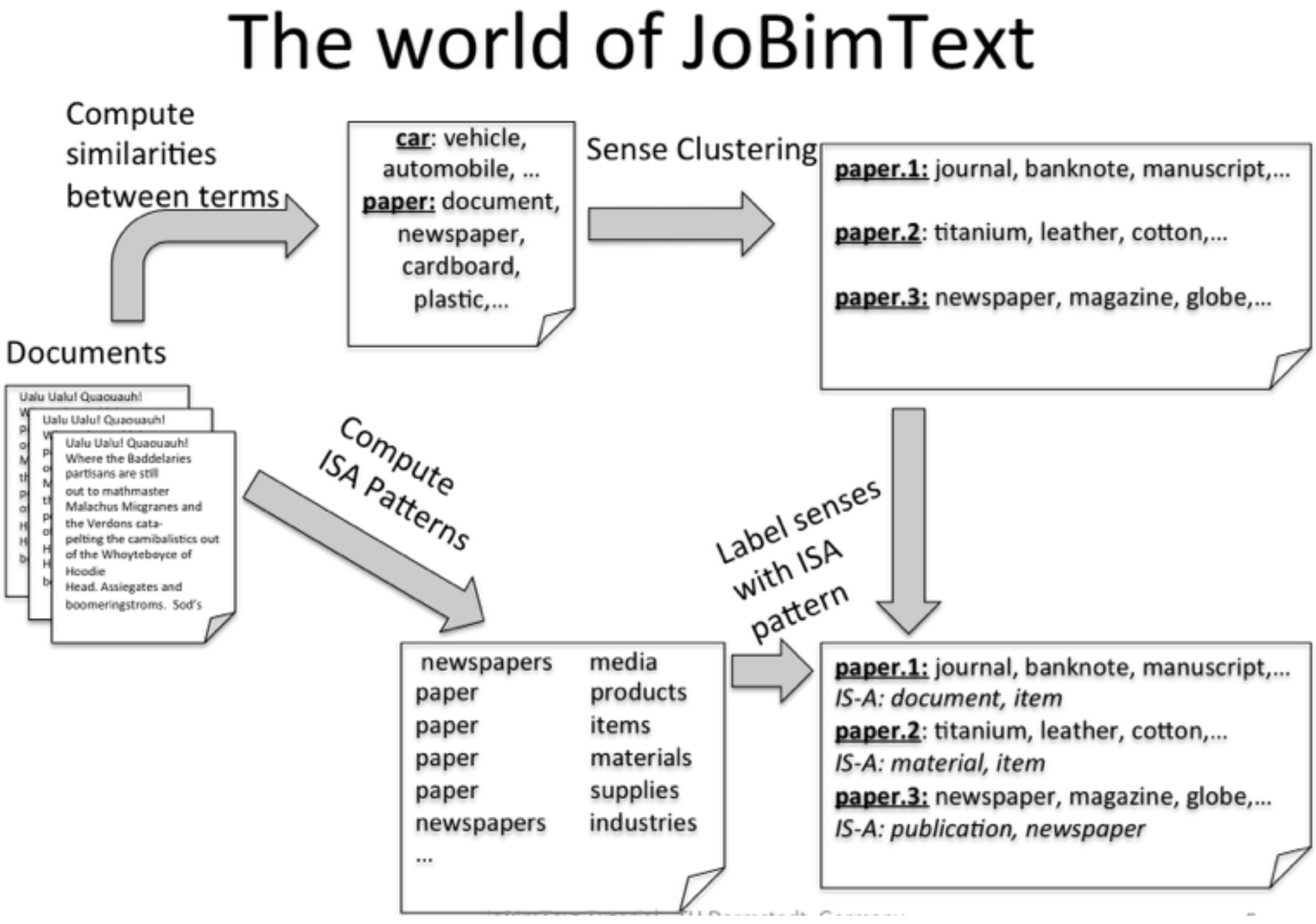
Language Element 1	Language Element 2	Score
cheese#n	stone#n	90
hard#a	yellow#a	50
...	...	...

Word Sense Induction

Sample word senses

Sense	Hypernyms	Similar lexical items	Aggregated Context Clues
beetle.0	car, company, macho nameplate, nameplate, icon, hit	camaro, mustang, gto, corvette, convertible, oldsmobil, minivan, camry, corolla, vw, impala, gt, thunderbird, jetta, convertible, gti, passat, sedan	<nn:car <nn:model <nn:dealership <nn:brand <nsubj:sell <dobj:drive <nsubj:have <nn:dealer <nn:owner <nn:vehicle <dobj:buy <nn:sale <nn:engine <nn:executive <nsubj:play >possessive:'s <nn:driver <nn:coupe <nsubj:offer <appos:car <dobj:own <nsubj:make <nsubj:announce <conj_and:bmw <poss:model <nn:convertible <nsubj:introduce >conj_and:bmw <nn:automobile <nsubj:car <nn:plant <nn:wagon <nn:engineer (…)
beetle.1	animal, species, insect, wildlife, creature	amphibian, bug, pythons, alligator, earwig, reptile, frog, bird, crocodile, wasp, grasshopper, earthworm, (.. 114 more) .., worm, butterfly, ladybug, parrot, gecko, cutworm, weevil, salamander, lemur	>det:the <dobj:kill <nsubj:are >det:these <dobj:find <nsubjpass:find >conj_and:insect >det:some <dobj:eat >det:a <prep_of:rid <nsubj:feed <dobj:keep <prep_of:species <dobj:call <nsubj:spread >amod:tiny <dobj:see <prep_of:type <conj_and:insect <prep_of:presence >det:those <prep_with:infested >cop:are <dobj:control <prep_of:number <dobj:remove >predet:all >conj_and:bug <nsubj:cause <nn:infestation <dobj:have <prep_from:protect >det:any >conj_and:beetle <prep_on:feed <conj_and:bug <prep_of:infestation <prep_of:spread >rcmod:feed <conj_and:flies <nn:problem <dobj:eliminate >partmod:find <dobj:prevent (…)
jaguar.0	car, brand, company, automaker, manufacturer, vehicle	geely, lincoln-mercury, tesla, peugeo, hyundai, nissan, kia, caravan, volvo, dodge, lexus, dealership, cruiser, corolla, daimlerchrysler, nike, cab, deville, vauxhall, rolls-royce, roadster, s.u.v., (.. 156 more ..) , jag, benz, mclaren, skoda, infiniti, sable, thunderbird	<dobj:drive >det:a >partmod:drive <dobj:buy >poss:his >partmod:park >amod:black <prep_of:driver >amod:new <dobj:driv >amod:white >det:the <appos:car <prep_of:seat >amod:blue >amod:red <dobj:own <prep_of:wheel >poss:her#PRP$ <prep_in:drive >amod:green <conj_and:car <prep_with:collide <nsubj:pull >rcmod:drive <prep_in:passenger >amod:silver >conj_and:car >amod:old <nsubj:crash <nsubjpass:steal <dobj:steal <nn:model <nsubj:strike <prep_in:ride <prep_of:back >conj_and:vehicle <nn:driver <prep_of:owner <dobj:sell >det:an <dobj:stop <prep_of:control <prep_into:get <agent:strike <nn:car <nsubj:collide >amod:gray >amod:steal <nsubjpass:park <nn:owner <nsubj:drive <conj_and:vehicle <nsubj:hit (…)
jaguar.1	animal, species, wildlife, team, wild animal, cat	panther, cougar, alligator, tiger, elephant, bull, hippo, dragon, leopard, shark, bear, otter, lynx, lion	>det:the <dobj:see >det:a <dobj:kill >amod:dead >amod:male <nn:population >nn:baby >amod:young >amod:female <nsubj:are <nsubjpass:find >conj_and:bird >amod:wild <nsubj:eat >conj_and:animal <prep_of:population <dobj:spot >amod:endangered <dobj:find >num:two <prep_of:number >amod:adult >amod:rare >amod:endangered >partmod:name >conj_and:species <prep_of:species >amod:stuffed >amod:giant <nsubj:species <prep_like:look <dobj:include >amod:large <conj_and:bears (…)

Induction of word senses from text



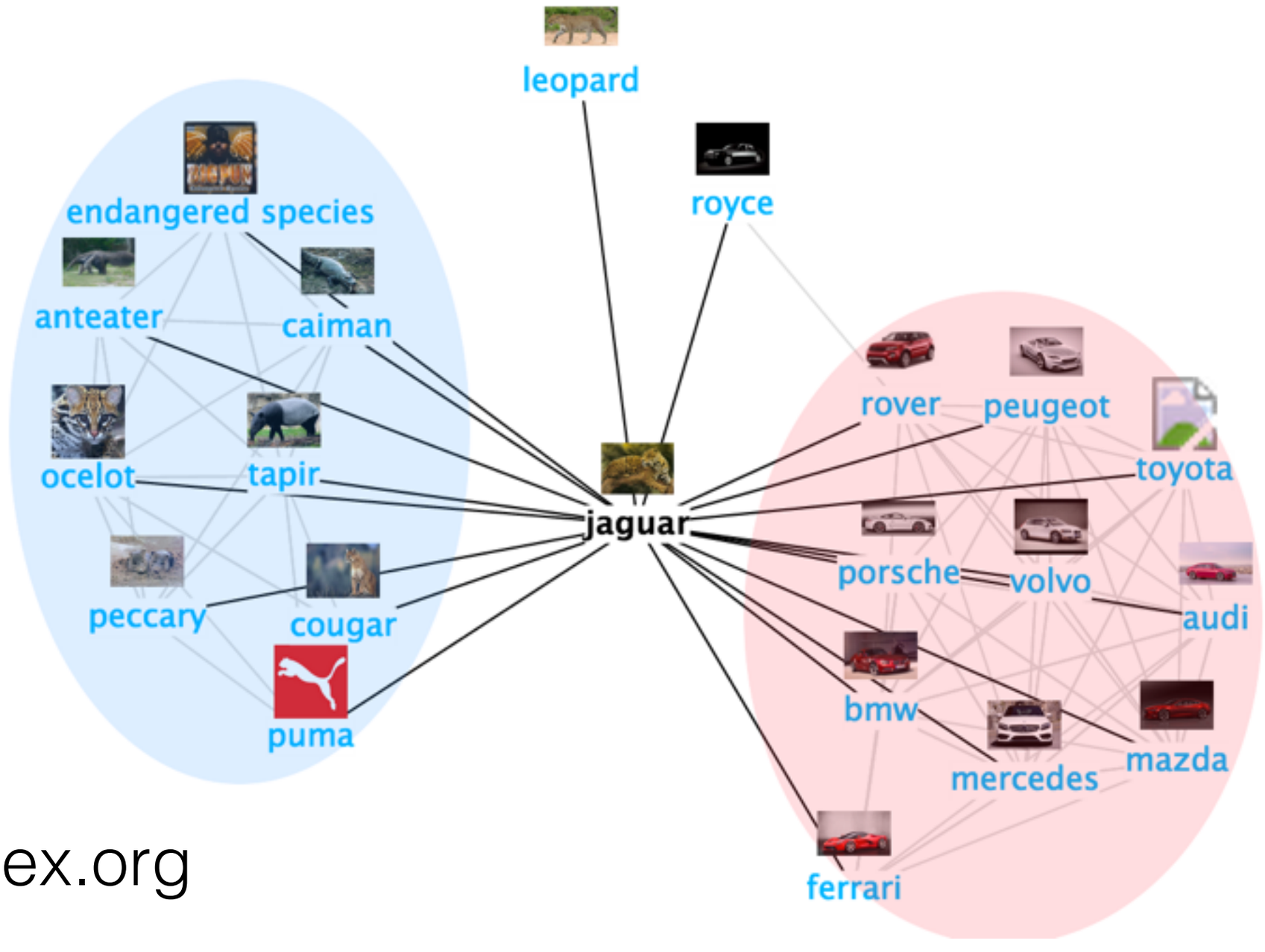
Mining word senses with ego network clustering

Word sense — a word cluster

Results count: 685

1 tapir
2 ocelot
3 puma
4 porsche
5 anteater
6 audi
7 cougar
8 mazda
9 rover
10 bmw
11 volvo
12 caiman
13 endangered species
14 ferrari
15 peugeot
16 toyota
17 leopard
18 mercedes
19 peccary
20 royce

Show next 20 results

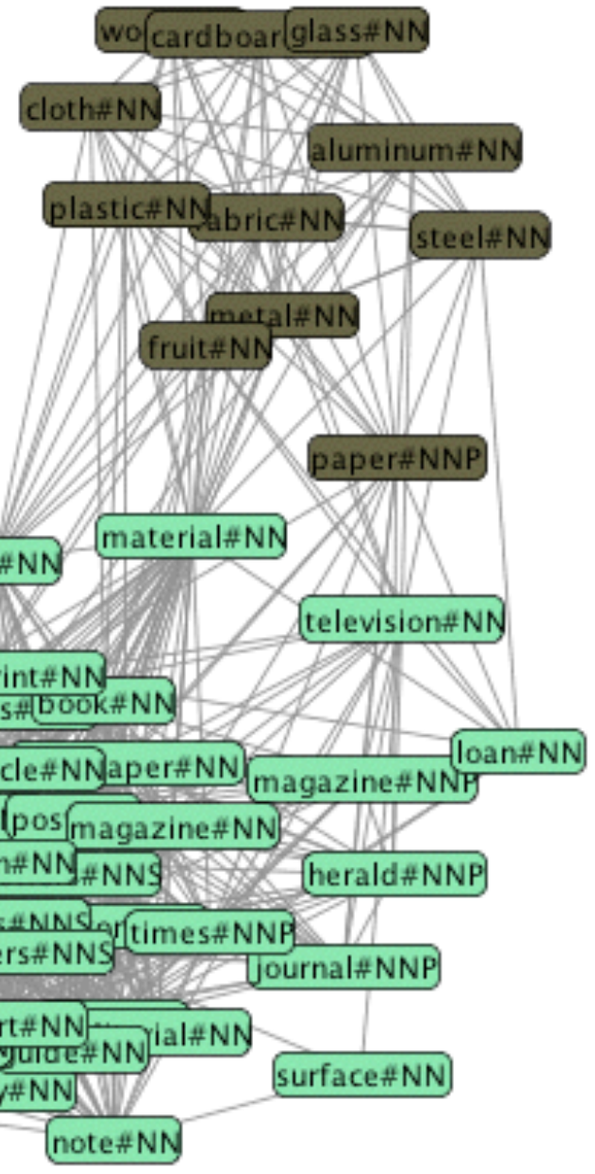


http://www.serelex.org

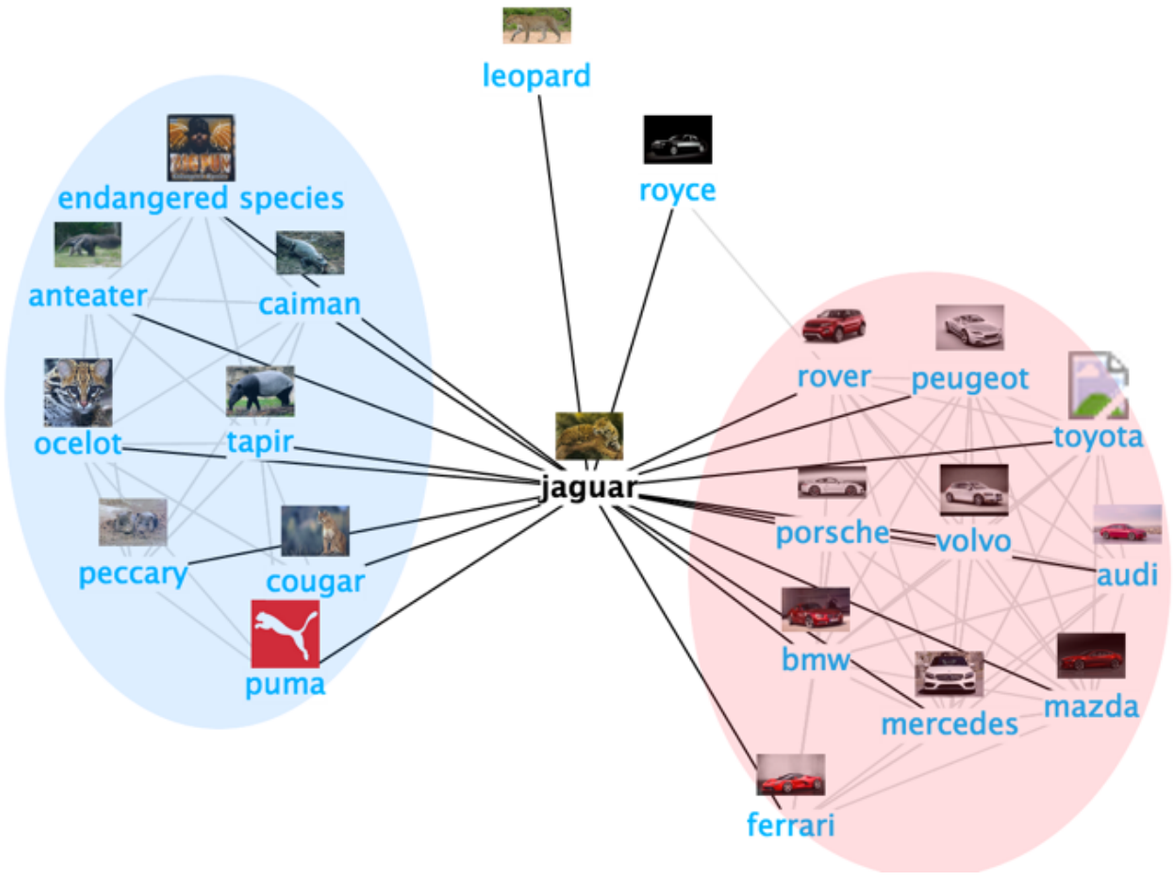
Mining word senses with ego network clustering

bar#NN

paper#NN



Hypernyms of word senses



Sense hypernyms — frequent IS-A relations in a word cluster

IS-A relations (~hypernyms)

- puma **is-a** {animal, cat}
  cougar **is-a** {animal, cat, speices}
- bmw **is-a** {car, brand, company}
  toyota **is-a** {car, company}

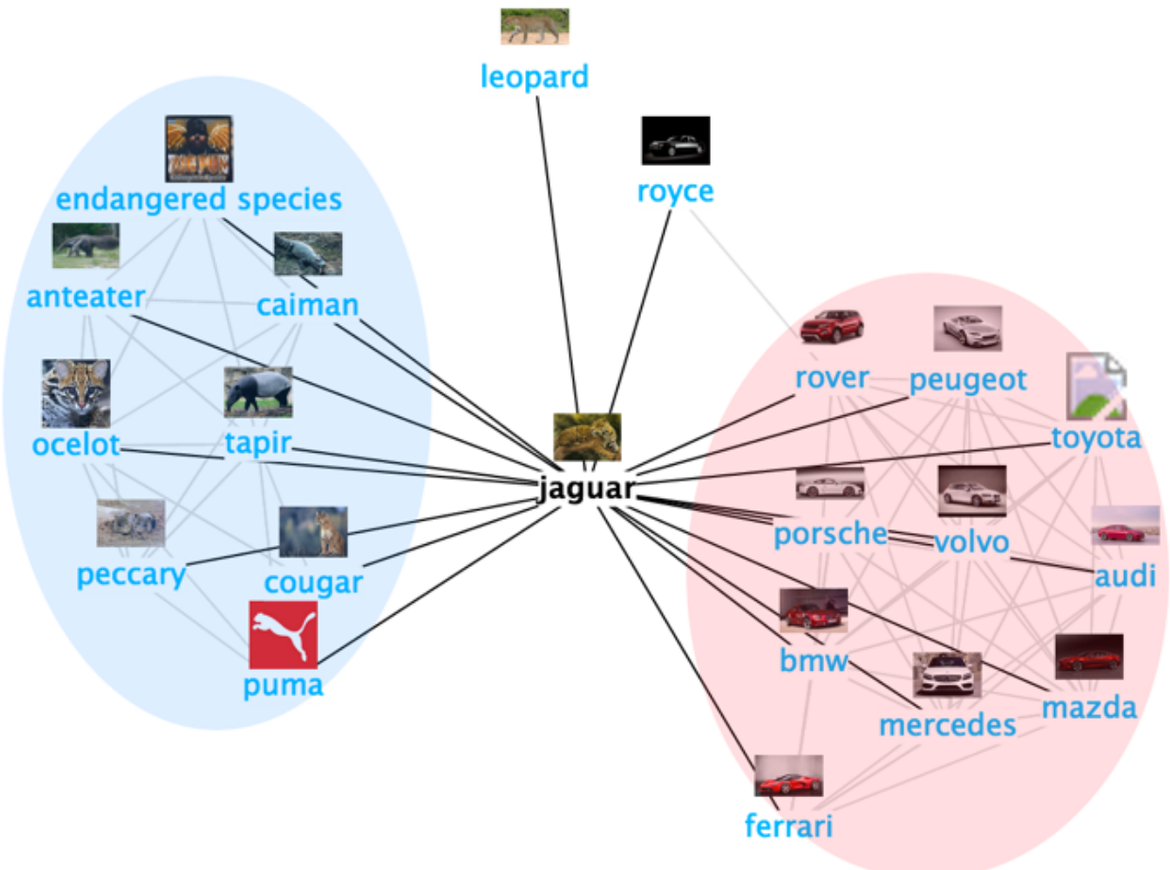
Hearst patterns

- 1. such NP as NP, NP[,] and/or NP;
- 2. NP such as NP, NP[,] and/or NP;
- 3. NP, NP [,] or other NP;
- 4. NP, NP [,] and other NP;
- 5. NP, including NP, NP [,] and/or NP;

Matches in text

- such {non-alcoholic [sodas=hyper]} as {[root beer=hypo]} and {[cream soda=hypo]}
- {traditional[food=hyper]}, such as {[sandwich=hypo]}, {[burger=hypo]}, and {[fry=hypo]}

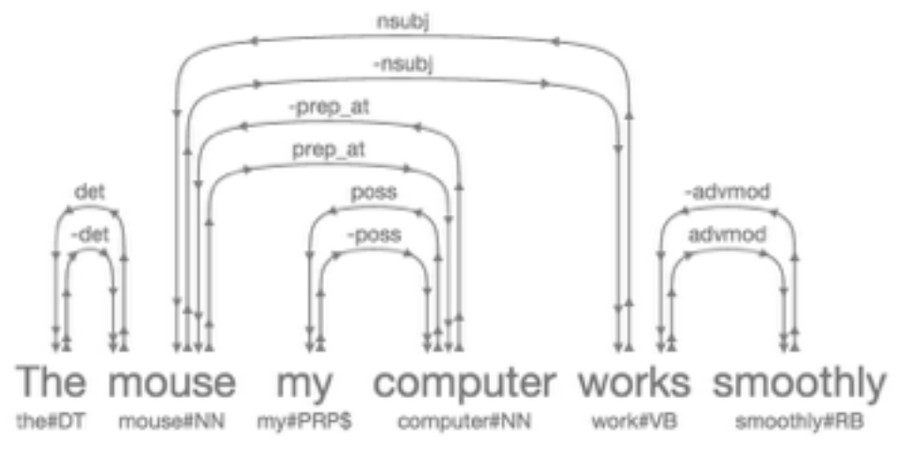
Context clues of word senses



Porsche	Corvette	Leopard	Lion
Bim	Bim	Bim	Bim
a#DT#det	his#PRP$#poss	snow#NN#nn	share#NN#-poss
drive#VB#-dobj	red#JJ#amod	a#DT#det	the#DT#det
his#PRP$#poss	chevrolet#NP#nn	lions#NN#-conj_and	a#DT#det
buy#VB#-dobj	a#DT#det	tigers#NN#-conj_and	mountain#NN#nn
maker#NN#nn	the#DT#det	chang#VB#-nsubj	sea#NN#nn
the#DT#det	white#JJ#amod	cloud#JJ#amod	den#NN#-poss
car#NN#nn	new#JJ#amod	the#DT#det	cub#NN#-nn
new#JJ#amod	million-plus#JJ#num	elephant#NN#-conj_and	dance#NN#-nn
red#JJ#amod	chevrolet#NN#nn	tiger#NN#-conj_and	heart#NN#-prep_of
driv#VB#-dobj	black#JJ#amod	lion#NN#-conj_and	tiger#NN#conj_and
black#JJ#amod	cheonan#NP#nn	spotted#JJ#amod	tamarin#NN#-nn
steal#VB#-dobj	sinkin#NN#-prep_of	amur#NP#nn	population#NN#-nn
no.#NN#nn	sixth-generation#JJ#amod	zebra#NN#conj_and	tamer#NN#-nn
luxury#NN#nn	frigate#NN#-conj_and	cat#NN#conj_and	aslan#NP#-appos
crash#VB#-dobj	drive#VB#-dobj	tigers#NN#conj_and	come#VB#-prep_like
drive#VB#partmod	2005#CD#num	tiger#NN#conj_and	head#NN#-poss
my#PRP$#poss	car#NN#conj_and	bears#NN#-conj_and	dancer#NN#-nn
ferrari#NP#-conj_and	navy#NN#nn	cheetah#NN#-conj_and	brigade#NN#prep_from

Context clues of a sense — frequent context features in a word cluster

JoBimText.org —> Web Demo



mouse#NN Jo

Count: 62650

Jos

Jo	Score
mouse#NN	746
rat#NN	192
rodent#NN	122
monkey#NN	112
pig#NN	103
animal#NN	95
human#NN	94

Bims

Bim	Score	Count
click#NN#-prep_of	14433.61	
a#DT#det	11612.08	
click#NN#-nn	9071.84	
the#DT#det	8613.77	
keyboard#NN#-conj_and	7548.80	
cat#NN#-conj_and	5417.09	
computer#NN#nn	4776.27	

CW

Sense 0 168: rat#NN · rodent#NN · mon…

Sense 1 32: keyboard#NN · joystick#NN ·…

Word Sense Disambiguation

Word Sense Disambiguation a.k.a. Contextualization

- **Goal:** use word sense inventory and apply it to text; assign the correct word sense based on the given context.

- **Example**: “python is a programming language with a great community”

Example of disambiguation w.r.t. word senses

python is a programming language with a great community

- **python5** [Python, JavaScript, perl, Perl, Fortran, …]
  hyper [language, languages, programming_language, programming_languages, scripting_language, technology, …]

- **is-1**

- **a-1**

- **programming0** [scripting, markup, Romance, Austronesian, spoken, Slavic, …]
  hyper [forms, groups, people, topics, …]

- **with2** [featured, featuring, included, includes, …]
  hyper []

- **a0** [some, two, several, many, …]
  hyper []

- **great0** [considerable, tremendous, huge, greater, immense, …]
  hyper [item, items]

- **community-1**

Example of disambiguation w.r.t. word senses

python snake is very dangerous

- **python5** [python4 [pythons, snake, cobra, rat, monster, viper, crocodile, …]
  hyper [animals, animal, species, specie, wildlife, creature, …]

- **snake0** [snakes, scorpion, cobra, spider, dragon, serpent, …]
  hyper [animals, animal, species, specie, …]

- **is-1**

- **very0** [extremely, fairly, quite, relatively, particularly, …]
  hyper []

- **dangerous0** [difficult, hazardous, powerful, deadly, challenging, …]
  hyper []

Disambiguation: Example

Mouse0	Mouse1	Mouse2	Mouse3
finger	rodent	software	malignant
thumb	guy	circuitry	embryonic
brain	baboon	users	fetal
skin	horse	screen	cancerous

Contextualization

```
Input: sentence, target words, proto-ontology
Output: senses for target words

for targetWord in sentence:
   originalBim = getBim(targetWord)
   similarBims = getSimilarBims(bim)
   for senseCluster in senseClusters(targetWord):
      for clusterTerm in senseCluster:
         for bim in {originalBim, similarBims}:
          if clusterTerm has bim: addScore(senseCluster)
   assignedSense = maxScore(senseClusters)
return { (targetWord, assignedSense) }
```

Thank you!Questions?