

Corpus-based Semantic Lexicon Induction with Web-based Corroboration

Sean Igo and Ellen Riloff

University of Utah

NAACL HLT 2009 UMSLLS Workshop

June 2009



The University of Utah
Biomedical Informatics

SCHOOL OF COMPUTING
THE UNIVERSITY OF UTAH 

Semantic Lexicon Induction

- Semantic lexicon:
 - Collection of words belonging to a semantic category
 - e.g. “bird” is an ANIMAL and “truck” is a VEHICLE
- Useful in many NLP tasks
 - Question answering
 - Information extraction
 - etc.
- Tedious / incomplete to compile by hand
- Induction = automation



Web-based Methods

- e.g., Paşca04, KNOWITALL
- Often meant to induce open-domain resources like WordNet.
- Rely on high-precision extraction patterns
 - strongly associate a word with a class
 - <class name> such as <class member> and *
- Ideally these lexicons would be suitable
- Web-based methods can overlook small, specialized domain-specific regions



Corpus-Based Methods

- Typically designed to induce domain-specific lexicons from a domain-specific corpus.
- Can find rare or non-intuitive terms
 - spelling variants (“tularemia”, “tularaemia”)
 - abbreviations, acronyms
 - jargon
- Useful for domain-oriented tasks like IE



Corroboration

- Corpus-based method proposes words
- Web search verifies class membership
 - with no further training necessary
- Best of both techniques
- Corpus-based supplies direction, Web supplies coverage



Basilisk: Bootstrapping lexicon learner

- Unannotated corpus
- Small set of *seed words*
 - chosen by human expert
- Complete set of *extraction patterns* (contexts)
- Basilisk finds contexts of seed words
- Proposes other words in those contexts as candidates
- Verifies that candidates occur in multiple likely contexts



Basilisk Extraction Patterns

- Lexico-syntactic context
- General purpose, not hyponym or class-related
- For example: Extract subject of verb “kidnap” in active voice

CF:

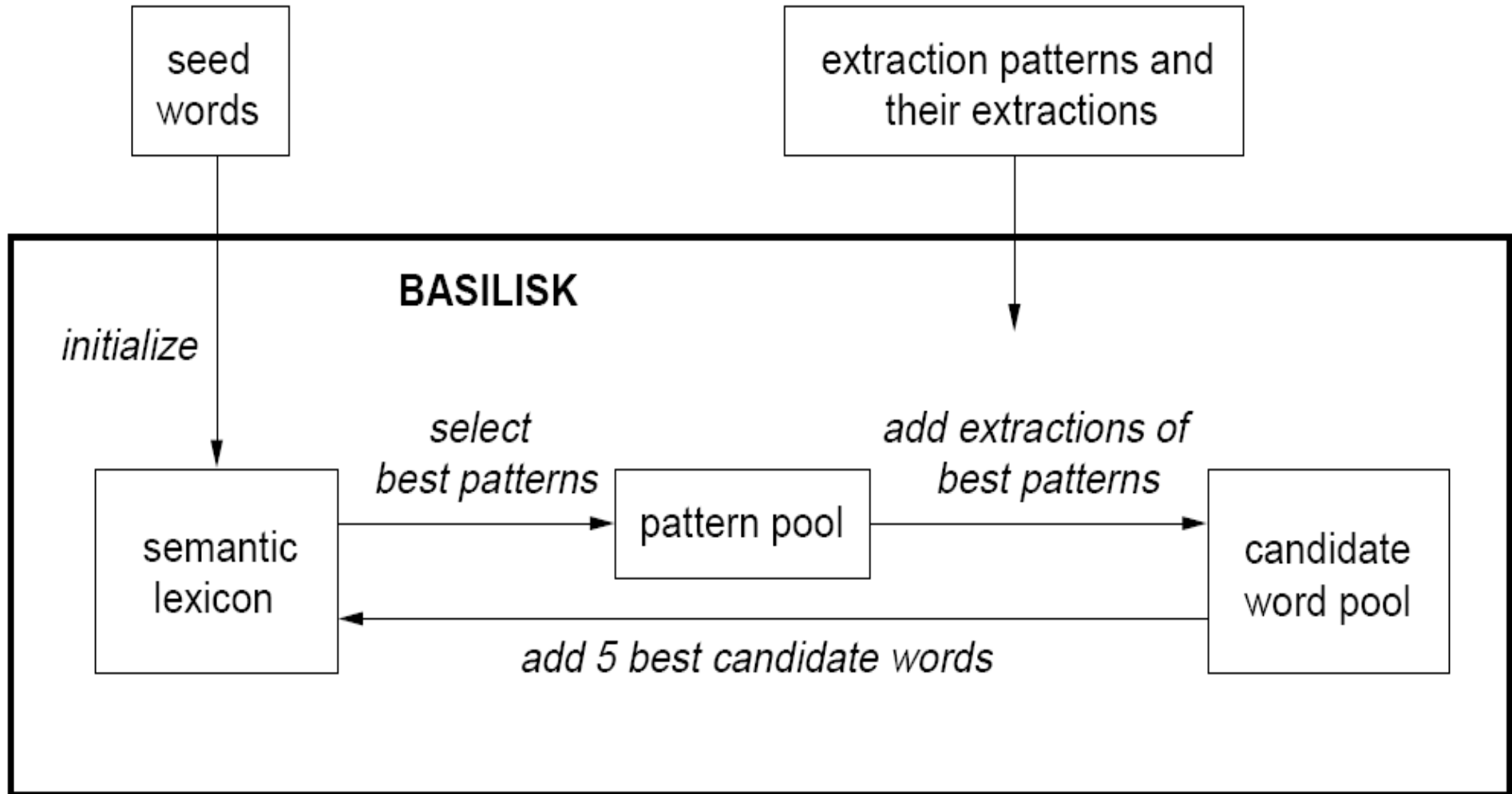
Name: <subj>_ActVp__KIDNAP_2329

Anchor: VP1 (KIDNAP)

Act_Fcns: active_verb_broad_p (VP1 (KIDNAP))

Slot: subj

Basilisk: iterative model



Improving Basilisk with web metrics

salmonell[osis
adenomatosis
under-investment
nationwide
bubonic_plague
theileria
bigemina
huge
kala-azar
rare
ross_river_virus
disinfecting
brucellosis
fa
hps
anthrax_disease
salmonellosis
litigation
trypanosomiasis
legionnaires'_disease
tick-borne_relapsing_fever
primary_pneumonic_plague
o139
anomalies
nv-cjd

- Basilisk subject to semantic drift
- Use web-based statistics to corroborate word pool choices:
 - Score candidates with web-based metric
 - Sort candidate words according to it
 - Choose top n or those scoring over a certain threshold
- Goal is to achieve a good reranking
 - Concentrate correct candidates at top
 - Could reorient Basilisk every so often



Web-based reranking metric

- Cf. Turney's PMI-IR with AltaVista NEAR
- Pointwise Mutual Information - $PMI(x,y)$

$$\log(\text{count}(x,y) / (\text{count}(x) \text{count}(y)))$$

- $\text{count}(x)$ is # AltaVista hits for term x
- $\text{count}(x,y)$ is # hits for x NEAR y
- -99999 score for not-found

Hypernym association

- Single class, not a hierarchy
- PMIs for word NEAR hypernym
 - e.g., PMI(“cow”, “animal”)
- Intuition: find definition-like patterns
 - “A *cow* is an *animal* found on farms...”
 - “An *animal* such as a *cow* has...”
- ...maybe no single hypernym always a good fit
 - “mosquito” / “animal”



Seed association

- PMIs for word wrt each seed
- Intuition: Find semantically similar items near one another (e.g., in lists)
- For example, with ANIMAL:
 - Seeds are *bird, mosquito, cow, horse, pig, chicken, sheep, dog, deer, fish*
- If candidate word is “rabbit” -
 - PMI(“rabbit”, “bird”); PMI(“rabbit”, “mosquito”); PMI(“rabbit”, “cow”) ...

Seed association

- Two different metrics
- Average of Seeds:
 - *average* of all PMI(word, seed_i) scores
 - Rewards word often found in context of category words
 - Suffers in cases where association is “specialized”
- Max of Seeds:
 - *maximum* of PMI(word, seed_i) scores
 - Rewards association with one good example
 - Susceptible to single over-strong entry



Method comparison - ANIMAL

Basilisk	Hypernym	Seed PMI Avg	Seed PMI Max
charcoal	ruminants	1.vi	1.vi
flanders	poultry	chickens	llangowan
ayam	swine	bird-to-bird	bird-to-bird
proteoglycans	raccoon	poultry	cervids
tse-tse	piglets	goats	goats
andalucia	mammals	pigs	ewes
navarra	rabbits	swine	ruminants
raccoon	dogs	rabbits	swine
leech	cats	geese	calf
goats	pigs	horses	lambs
phoebes	opossum	cows	wolsington
bison	heifer	dogs	piglets
czech_republic	bird-to-bird	raccoon	heifer
girl	puppy	bison	flocks
macaque	rabid	mammals	elk
bpsmith@ucdavis.edu	cheetah	cats	cows
buck	chickens	calf	pigs
lipophosphoglycan	herds	ducks	herds
melnick	elephants	ox	guinea
bear	goats	elk	canidae



Experiments

- 2 domains, 7 categories in all
- MUC-4 Latin American terrorism corpus
 - WEAPON
 - LOCATION
 - HUMAN
 - BUILDING
- ProMED disease outbreak corpus
 - ANIMAL
 - DISEASE
 - SYMPTOM



Gold standard

visit event
evidence none
students human
aggression event
candidate human
ambassador human
details none
mexico location
car vehicle
explosives weapon
names none
official human
station building
bolivia location

- Human experts created a dictionary
- Associates words with the best semantic category
- Domain-specific
- One category per word



Experiments

- For each category:
- Human expert chooses 10 seed words
- Basilisk generates 300 candidate words
- Build a profile of how accurate its ordering is
 - accuracy of first 25 words
 - accuracy of first 50 words
 - ...
 - accuracy of first 100 words
- Concentration of correct words over iterations



Scoring web-based corroboration

- Goal was to rerank Basilisk words
- Concentrate correct words at top
- Scored our system for accuracy
 - accuracy of first 25 words
 - accuracy of first 50 words
 - ...
 - accuracy of first 100 words
- Compare methods with each other and Basilisk

Scoring Example

ANIMAL				
N	Basilisk			
25	12/25			
300				

Scoring Example

ANIMAL				
N	Basilisk			
25	0.48			
300				

Scoring Example

ANIMAL				
N	Basilisk			
25	0.48			
50	0.58			
75	0.55			
100	0.45			
300				

Scoring Example

ANIMAL				
N	Basilisk	Hypernym		
25	0.48	0.88		
50	0.58	0.82		
75	0.55	0.68		
100	0.45	0.55		
300				

Scoring Example

ANIMAL				
N	Basilisk	Hypernym	Seed Avg	
25	0.48	0.88	0.92	
50	0.58	0.82	0.84	
75	0.55	0.68	0.67	
100	0.45	0.55	0.54	
300				

Scoring Example

ANIMAL				
N	Basilisk	Hypernym	Seed Avg	Seed Max
25	0.48	0.88	0.92	0.92
50	0.58	0.82	0.84	0.80
75	0.55	0.68	0.67	0.69
100	0.45	0.55	0.54	0.57
300				

Scoring Example

ANIMAL				
N	Basilisk	Hypernym	Seed Avg	Seed Max
25	0.48	0.88	0.92	0.92
50	0.58	0.82	0.84	0.80
75	0.55	0.68	0.67	0.69
100	0.45	0.55	0.54	0.57
300				

Scoring Example

ANIMAL				
N	Basilisk	Hypernym	Seed Avg	Seed Max
25	0.48	0.88	0.92	0.92
50	0.58	0.82	0.84	0.80
75	0.55	0.68	0.67	0.69
100	0.45	0.55	0.54	0.57
300	0.20	0.20	0.20	0.20

Scores: MUC-4 domain

<i>N</i>	BUILDING				HUMAN				LOCATION				WEAPON			
	<i>Ba</i>	<i>Hy</i>	<i>Av</i>	<i>Mx</i>	<i>Ba</i>	<i>Hy</i>	<i>Av</i>	<i>Mx</i>	<i>Ba</i>	<i>Hy</i>	<i>Av</i>	<i>Mx</i>	<i>Ba</i>	<i>Hy</i>	<i>Av</i>	<i>Mx</i>
25	.40	.56	.52	.56	.40	.72	.80	.84	.68	.88	.88	1.0	.56	.84	1.0	1.0
50	.44	.56	.46	.40	.56	.80	.88	.86	.80	.86	.84	.98	.52	.74	.76	.90
75	.44	.45	.41	.39	.65	.84	.85	.85	.80	.88	.80	.99	.52	.63	.65	.79
100	.42	.41	.38	.36	.69	.81	.80	.87	.81	.85	.78	.95	.55	.55	.56	.63
300	.22				.82				.75				.26			

Ranking results showing accuracies for the top-ranked *N* words.

Ba=Basilisk, *Hy*=Hypernym Re-ranking, *Av*=Average of Seeds Re-ranking, *Mx*=Max of Seeds Re-ranking



Scores: ProMED domain

<i>N</i>	ANIMAL				DISEASE				SYMPTOM			
	<i>Ba</i>	<i>Hy</i>	<i>Av</i>	<i>Mx</i>	<i>Ba</i>	<i>Hy</i>	<i>Av</i>	<i>Mx</i>	<i>Ba</i>	<i>Hy</i>	<i>Av</i>	<i>Mx</i>
25	.48	.88	.92	.92	.64	.84	.80	.84	.64	.84	.92	.80
50	.58	.82	.84	.80	.72	.84	.60	.82	.62	.76	.90	.74
75	.55	.68	.67	.69	.69	.83	.59	.81	.61	.68	.79	.71
100	.45	.55	.54	.57	.69	.78	.58	.80	.59	.71	.77	.64
300	.20				.62				.38			

Ranking results showing accuracies for the top-ranked *N* words.

Ba=Basilisk, *Hy*=Hypernym Re-ranking, *Av*=Average of Seeds Re-ranking, *Mx*=Max of Seeds Re-ranking



Filtering

θ	Category	Acc	Cor/Tot
-22	WEAPON	.88	46/52
	LOCATION	.98	59/60
	HUMAN	.80	8/10
	BUILDING	.83	5/6
	ANIMAL	.91	30/33
	DISEASE	.82	64/78
	SYMPTOM	.65	64/99
-23	WEAPON	.79	59/75
	LOCATION	.96	82/85
	HUMAN	.85	23/27
	BUILDING	.71	12/17
	ANIMAL	.87	40/46
	DISEASE	.78	82/105
	SYMPTOM	.62	86/139
-24	WEAPON	.63	63/100
	LOCATION	.93	111/120
	HUMAN	.87	54/62
	BUILDING	.45	17/38
	ANIMAL	.75	47/63
	DISEASE	.74	94/127
	SYMPTOM	.60	100/166

Filtering results using Max of Seeds.

- Use a score threshold to disqualify candidates
- Too lenient lets in bad choices
- Too strict takes many iterations
- No single score seems suitable across categories

Conclusions

- Web co-occurrence statistics can improve the ranking of lexicon entries without requiring any additional supervision
- Max of Seeds works best of these methods
- Filtering doesn't work – with these metrics
- Future work:
 - try other metrics, such as median seed PMI
 - Incorporate into Basilisk bootstrapping



Questions?

