Growing Fine-Grained Concept Hierarchies

from Seeds of
Varying Quality and Size

Tony Veale and Co.

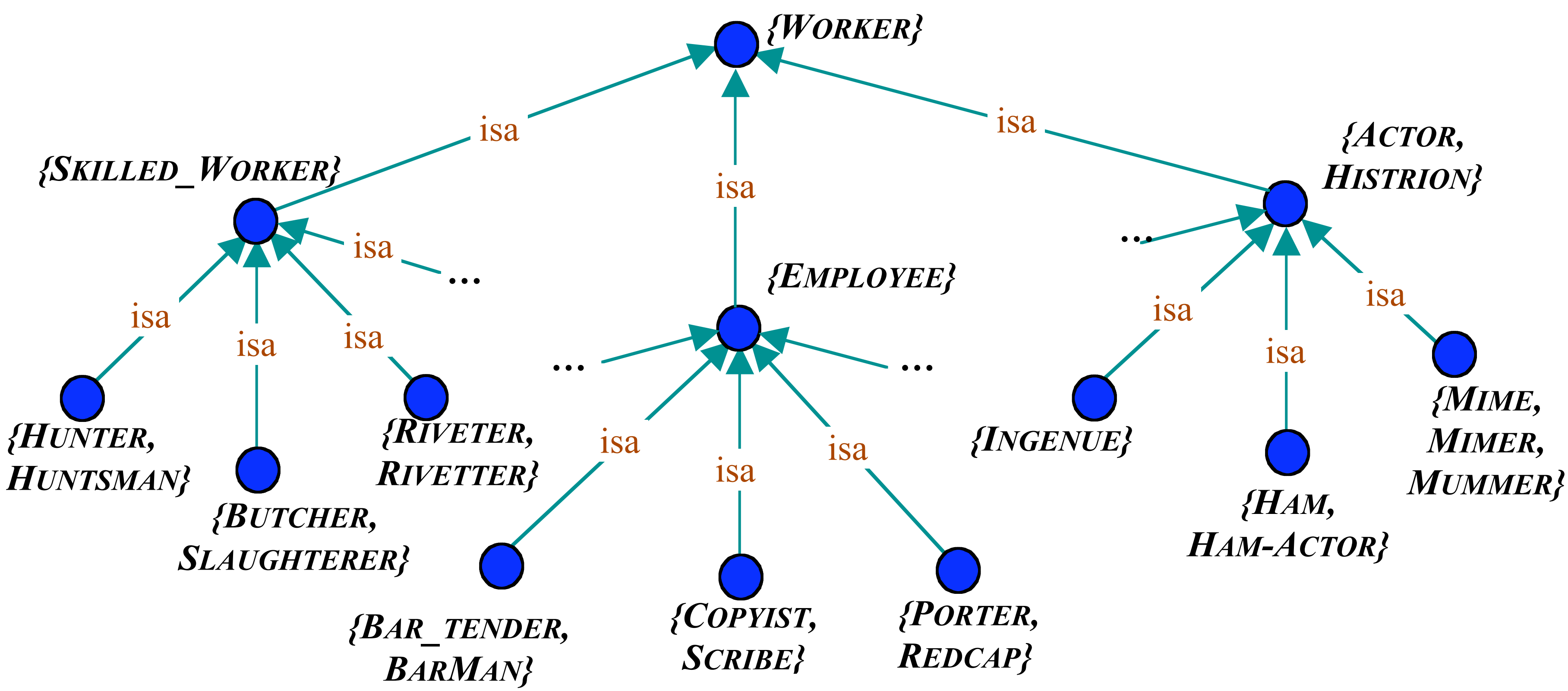
School of Computer Science
and informatics, UCD

Creative Language Systems Group

Tony.Veale@UCD.ie


Guofu Li
Cris Butnariu
Yanfen Hao

WordNet: An Linguistically-Motivated Concept Hierarchy



Synsets denote word-senses **But do synsets capture real conceptual categories?**

What Should a Term Hierarchy Provide by Way of Meaning?

WordNet is a "lightweight ontological" approach to lexical semantics

WN provides a deep(-ish) hierarchy for nouns, but lacks explicit "meaning"

- ***Aristotelian Taxonomies, Description Logics*** — **WordNet, *CYC*, SUMO**

  More Explicit, ***BUT*** tries to draw sharp lines between overlapping categories

- **Explicit Semantics / First-Order Logic** — ***SUMO*, HowNet, *CYC***

  Supports inference & theorem proving, ***BUT*** highly selective and often sparse

- **"Firthian" Corpus-Based Approaches** — **Firth, Sinclair, Hanks**

  Ecological sensitivity to word-usage, ***BUT*** lacks definitive ontology structure

Building Ontologies: **Different Needs, Different Approaches**

- Handcrafted, Knowledge-Engineering

  E.g., CYC, WordNet, SUMO, HowNet, etc.

- Conversion from authoritative Sources

  E.g., MRDs (Longman's LDOCE, etc.), Wikipedia

- Direct Extraction from Corpora using IE

  E.g., by looking for "X is a [kind of] Y" patterns

- Indirect Extraction from Corpora (via clustering)

  E.g., by acquiring diagnostic criteria, and clustering a taxonomy

- Bootstrapping from Corpora \ World-Wide-Web

  E.g., using a seed-base of existing knowledge to acquire more from text

The Knowledge Spectrum

- **Knowledge-Based Inference Systems *(e.g., CYC / CYC-ANSWERS)***

  Which countries will be capable of launching a spy satellite by 2010?

  - **Productive but expensive knowledge, expensive inference demands**

- **80/20 Techniques *(mapping standardized problems to procedural semantics)***

  Who is the CEO of IBM? ⇒ select CEO from Company where Name = ‘IBM’

- **Shallow Statistical Techniques and Information Retrieval**

  - **The contestants in TREC 1999 - 2001 Q&A (SMU Arrow/Falcon etc.)**
  - **Q&A = Shallow NLP + Information Retrieval + Information Extraction**
  - **Knowledge-Base = World-Wide-Web / Private Text Archive**
  - **Accidental Experts: Vast information reach but limited inference capability**

Cyc's Knowledge-Rich Ontology Supports Analogy

HPKB TQO125c **How is a terrorist group's interest in group cohesion like a criminal organization's interest in maintaining security?**

Answer:

Like criminal organizations, terrorist groups have an interest in keeping their membership cohesive to maintain their security. A fragmented and disloyal membership can compromise a group's safety, undermine its operations, and threaten its survival.

Source(s):

1. ***Organized Crime in the Former Soviet Union Fact sheet.***
2. ***International System Framework.***

Cyc's Knowledge-Base Also Supports Disanalogy

HPKB TQO125b **How is a terrorist group's interest in increasing financial assets different from a criminal organization's interest in earning profits?**

Answer:

1. ***Each group's interest reflects different goals.***
2. ***A terrorist group's interest in increasing its financial assets, while important, is not its main purpose. Rather, acquiring assets is the means by which the group meets its operational and organizational requirements and achieves its goals. A criminal organization's interest in earning profits, in contrast, is its central goal.***

Source(s):

1. ***Organized Crime in the Former Soviet Union Fact sheet.***

Top-Down Knowledge Engineering (KE) in Cyc's Ontology

Knowledge Engineering is a process of ***Ontologization*** and ***Axiomatization***

```
(#$forAll ?PER
  (#$thereExists ?FANCLUB
      (#$implies (#$and(#$isa ?PER FamousPerson)
                       (#$isa ?FANCLUB (#$MobFn Person))
                       (#$groupMembers ?FAN ?FANCLUB))
                 (#$feelsTowardsObject ?FAN ?PER
                                  #$Admiration #$Positive))))
```

Axioms are associated with concepts (collections or individuals) in **microtheories**.

Implication Axioms *(rules)* can be designated as ***forward***- or ***backward***- firing.

Rule-Bound Reasoning

- *At the Core of CYC is an* **Ontology** *of Concepts (Taxonomy + Relationships)* that informs and underpins all axioms in the KB.
- These concept representations do not reflect current thinking in the **cognitive psychology of category structure** *(e.g., radial, fuzzy, prototype-based).*

For Example, consider how Cyc combines concepts for Noun-Noun compounds:

```
(#$nnRule "potato gun"
    (#$and  (#$genls :NOUN1 #$PartiallyTangible)
            (#$genls :NOUN2 #$ProjectileLauncher)
            (#$not
                (#$genls :NOUN1 #$Organism-Whole)))
    (#$isa  :NOUN
            (#$SubcollectionOfWithRelationToTypeFn
                        :NOUN2 #$launchesProjectile :NOUN1))
```

But there are many problems with this account:

Concepts should combine as a matter of definition and meaning; rules are easily defeated and too top-down.

"Authoritative" Hand-Crafting leads to Over-Specification

- Excessive (and *obsessive*) **Ontologization** can lead to **hair-splitting**.
- For example, Cyc discriminates among many different senses of "**in**" :

E.g.,

in *(full submerged)*	**– like an olive *in* a martini**
in *(partially submerged)*	**– like a toothpick *in* the olive**
in *(surrounded by)*	**– like a man *in* a field**
in *(membership)*	**– like a man *in* a club**

But strangely, not:

in *(abstract situation)*	**– like a woman *in* love**
in *(content area)*	**– like an academic in a research field**

These copious *(and uneven)* discriminations yield a combinatorial explosion for NLP parsing systems, *yet fail to capture the true essence of "in".*

The Excluded Middle

- Cyc supports two Truth values: **True** and **False** *(no middle ground)*

- Cyc supports two Truth modalities:

  Default (*defeasible*) and **Monotonic** (*indefeasible*).

- Cyc does not represent facts probabilistically *(e.g., 80% likelihood)* or fuzzily.

  This makes it very difficult to axiomatize typical (but not analytic) truths, such as sandwiches comprise two pieces of bread with meat inside.

```
(#$typicalWRT  #$Penguin      #$ArcticBird)
(#$atypicalWRT #$Penguin      #$Bird)
(#$atypicalWRT #$Insect       #$Food)
(#$typicalWRT  #$Calzone      #$ItalianCuisine)
(#$atypicalWRT #$Calzone      #$Pizza)
```

Real common-sense informs us when a situation is **atypical**, **unexpected** or **surprising**. Without typicality, we are left with **possibility** *versus* **impossibility**.

Direct Extraction from Text: Using "Hearst (1992)" Patterns

<u>Singly-Anchored Retrieval Patterns</u>

One "anchor" term can be used to retrieve relations from the WWW

*			*		*
X *(s)*	**such**	**as**	**Y** ***+plural***	**and**	**Z** *(s)*

E.g., "<u>Antidepressants</u> such as Prozac and Zoloft"

*		*			*
Z ***+plural***	**and**	**X** ***+plural***	**and**	**other**	**Y** ***+plural***

E.g., "Worms, <u>Viruses</u> and other malicious programs"

Problem: ***These patterns are robust but relatively infrequent in most texts.***

Bottom-Up Approaches: Using the "Distributional Profile" of a term

- **Noun used as the *subject / object* of an active verb** **(Role Noun Verb)**

  E.g., a virus *infects*, a robot *obeys*, an opera is *composed*, etc.

- **Noun modified by a given adjective** **(Attr Noun Adj)**

  E.g., insults are *hurtful*, clichés are *tired*, priests are *religious*, etc.

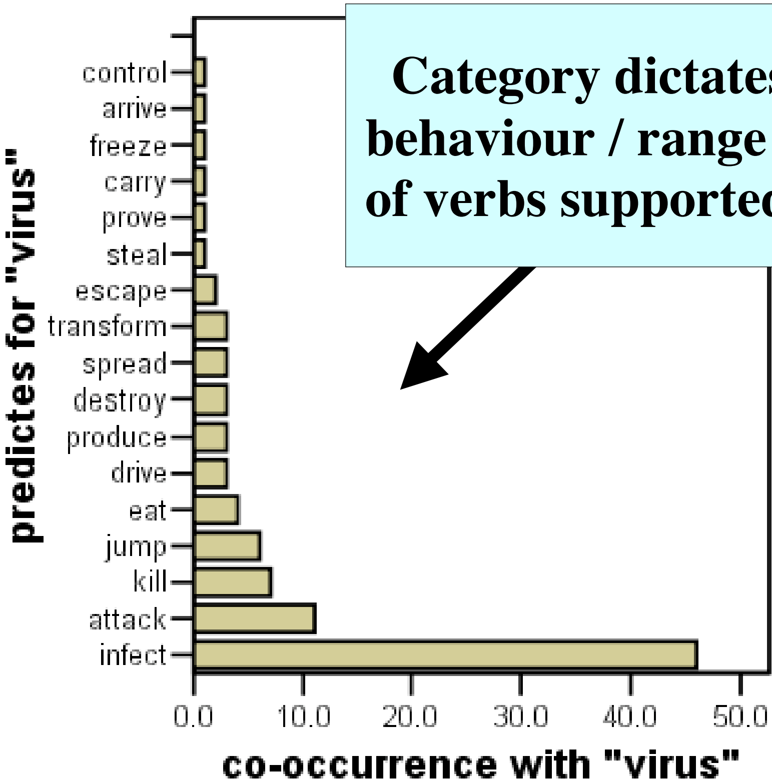
- **Noun used in a "Group of X" construction** **(Group Noun Noun)**

  E.g., an *army* of soldiers, a *conclave* of bishops, a *posse* of rappers, etc.

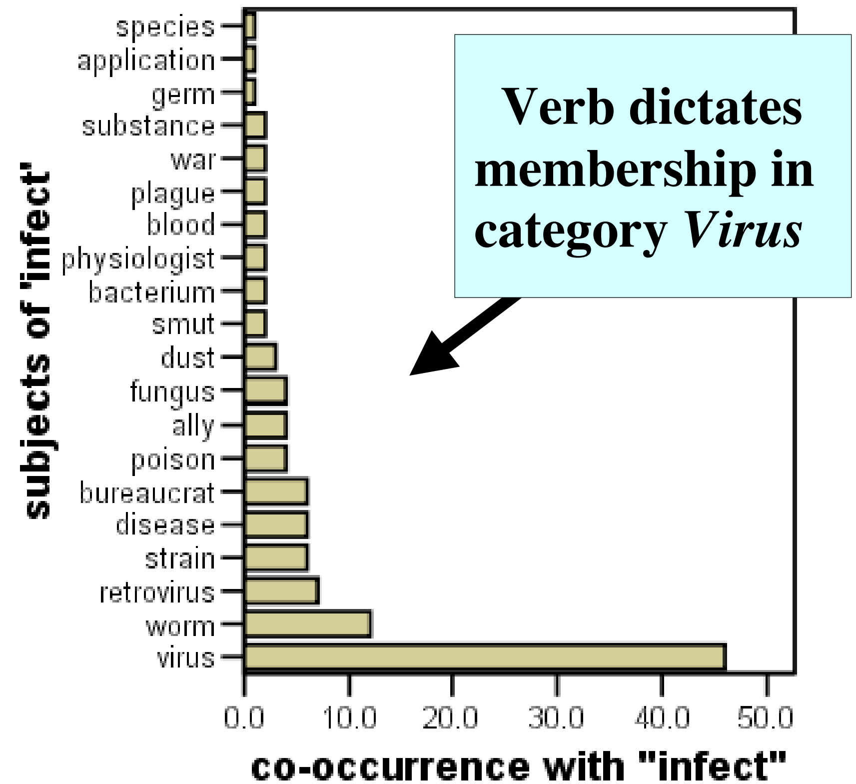
- **Noun used in a PP-phrase with a given prep. head** **(Attach P Noun)**

  E.g., *against* an adversary, *via* an intermediary, *along* a channel, etc.

Wikipedia as a Distributional Context: “Virus” and “Infect”



Category dictates behaviour / range of verbs supported



Verb dictates membership in category *Virus*

On the web: `afflatus.ucd.ie` *(current projects / Lex-Ecologist)*

Acquiring Qualia Structures from Textual Patterns on the WWW

Formal (IS-A)	*Constitutive (Made-Of)*
"an X is a kind of Y"	"an X is made up of Ys"
"an X is Y"	"an X is made of Ys"
"an X and other"	"an X comprises Ys"
"an X or other"	"an X consists of Ys"
"Ys such as Xs"	"Xs are made up of Ys"
"Xs and other Ys"	"Xs are made of Ys"
"Xs or other Ys"	"Xs comprise Ys"
"Ys, especially Xs"	"Xs consist of Ys"

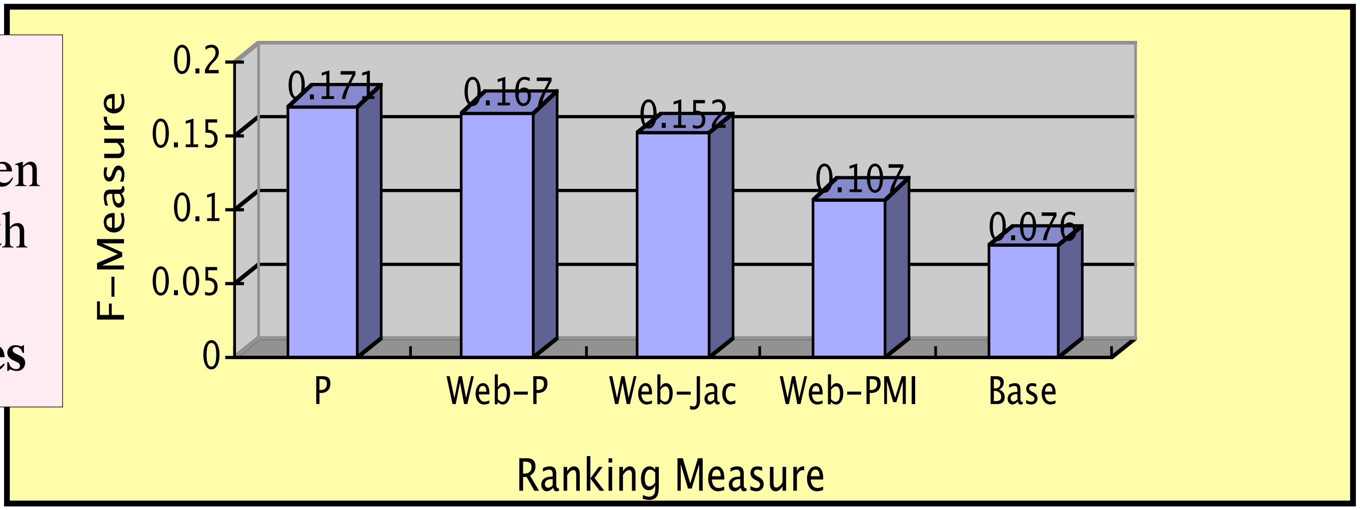
Cimiano, P.
Wenderoth, J.
ACL 2007

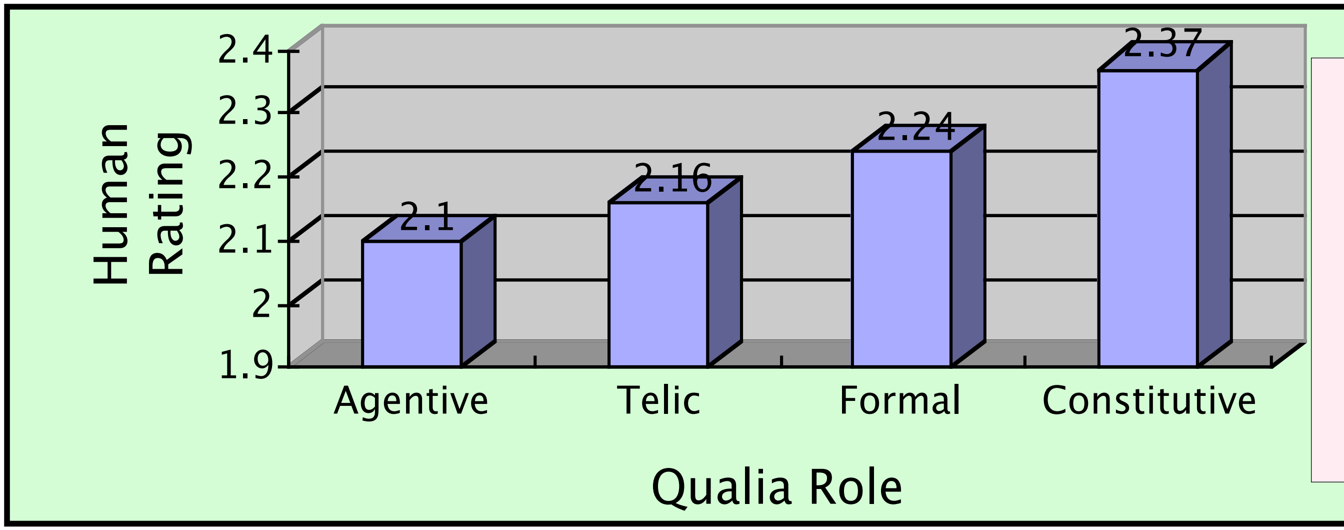
Telic (Used for)	*Agentive (is Made by)*
"purpose of an X is"	"to VERB a new x"
"an X is used to"	"to VERB a complete x"
"purpose of Xs is"	"a new X has been *Yed*"
"Xs are used to"	"complete X has been *Yed*"

Extracting Qualia: Empirical Results

Cimiano, P. and Wenderoth, J. (2007). Automatic Acquisition of Ranked Qualia Structures from the Web. *In Proc. of the 45th Annual Meeting of the ACL*, pp 888-895.

Very Low **F-scores** when compared with independent **human values**

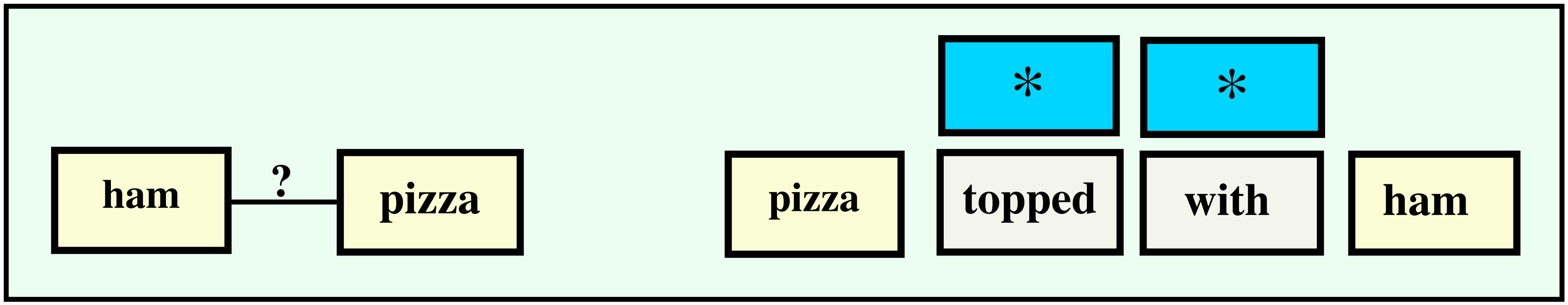




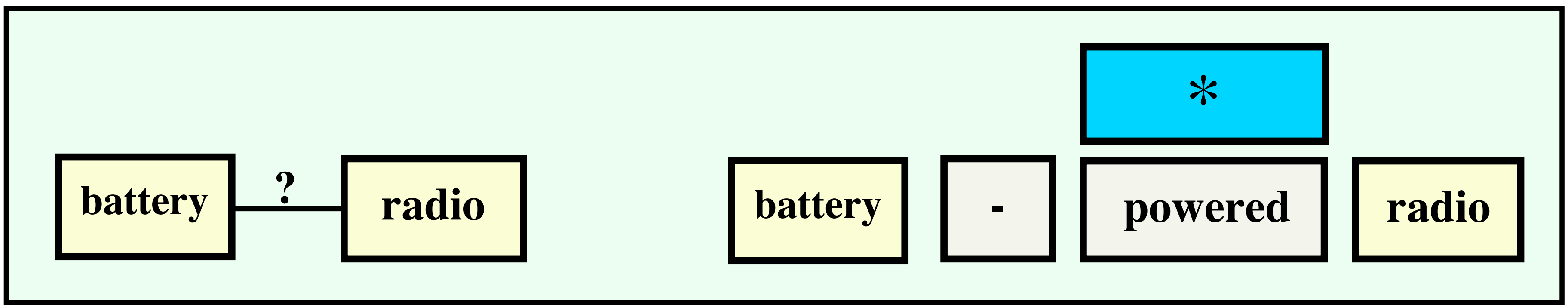
Quite **Good Evaluations (3 *highest*)** when assessed by humans **after-the-fact.**

Finding Relations Between Terms: Using WWW to "fill in the blanks"

Noun-Compounds (NCs) are a special case of compressed ontological relations



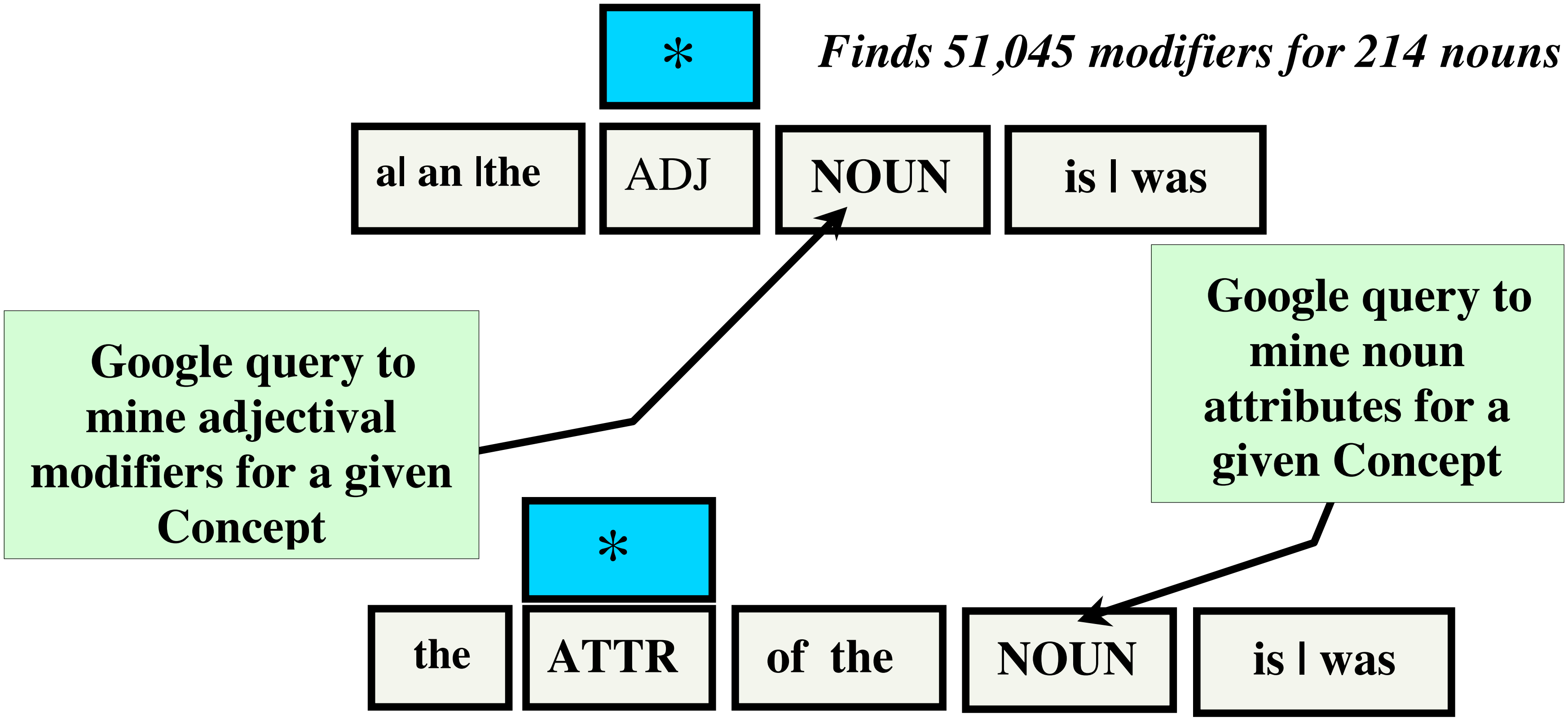
E.g., Nakov, Hearst, Turney, Butnariu and Veale, …



The WWW can be used as a corpus for finding <u>missing relations</u> between terms

Almuhareb+Poesio (2004): Web-Mining of Concept Modifiers/Attributes

Finds 51,045 modifiers for 214 nouns



Finds 8934 attributes for 214 nouns

e.g., rocket = [fast, powerful, speed, thrust, …] **vector space of 59,979 features**

Almuhareb+Poesio (2004): Clustering Concepts by Modifiers/Attributes

Class	Concepts
Animal	bear, bull, camel, cat, cow, deer, dog, elephant, horse, kitten, lion, monkey, mouse, oyster, puppy, rat, sheep, tiger, turtle, zebra
Building	abattoir, center, clubhouse, dormitory, greenhouse, hall, hospital, hotel, house, inn, library, nursery, restaurant, school, skyscraper, tavern, theater, villa, whorehouse
Cloth	pants, blouse, coat, costume, gloves, hat, jacket, jeans, neckpiece, pajamas, robe, scarf, shirt, suit, trousers, uniform
Creator	architect, artist, builder, constructor, craftsman, designer, developer, farmer, inventor, maker, manufacture, musician, originator, painter, photographer, producer, tailor
Disease	acne, anthrax, arthritis, asthma, cancer, cholera, cirrhosis, diabetes, eczema, flu, glaucoma, hepatitis, leukemia, malnutrition, meningitis, plague, rheumatism, smallpox
Feeling	anger, desire, fear, happiness, joy, love, pain, passion, pleasure, sadness, sensitivity, shame, wonder
Fruit	apple, banana, berry, cherry, grape, kiwi, lemon, mango, melon, olive, orange, peach, pear, pineapple, strawberry, watermelon
Furniture	bed, bookcase, cabinet, chair, couch, cradle, desk, dresser, lamp, lounge, seat, sofa, table, wardrobe
Body Part	ankle, arm, ear, eye, face, finger, foot, hand, head, leg, nose, shoulder, toe, tongue, tooth, wrist
Publication	atlas, book, booklet, brochure, catalog, cookbook, dictionary, encyclopedia, handbook, journal, magazine, manual, phonebook, reference, textbook, workbook
Family Relation	boy, child, cousin, daughter, father, girl, grandchild, grandfather, grandmother, husband, kid, mother, offspring, sibling, son, wife
Time	century, decade, era, evening, fall, hour, month, morning, night, overtime, quarter, season, semester, spring, summer, week, weekend, winter, year
Vehicle	aircraft, airplane, automobile, bicycle, boat, car, cruiser, helicopter, motorcycle, pickup, rocket, ship, truck, van

214 concepts from 13 PWN categories

402 concepts from 21 PWN categories

Almuhareb & Poesio (2004): Clustering Results

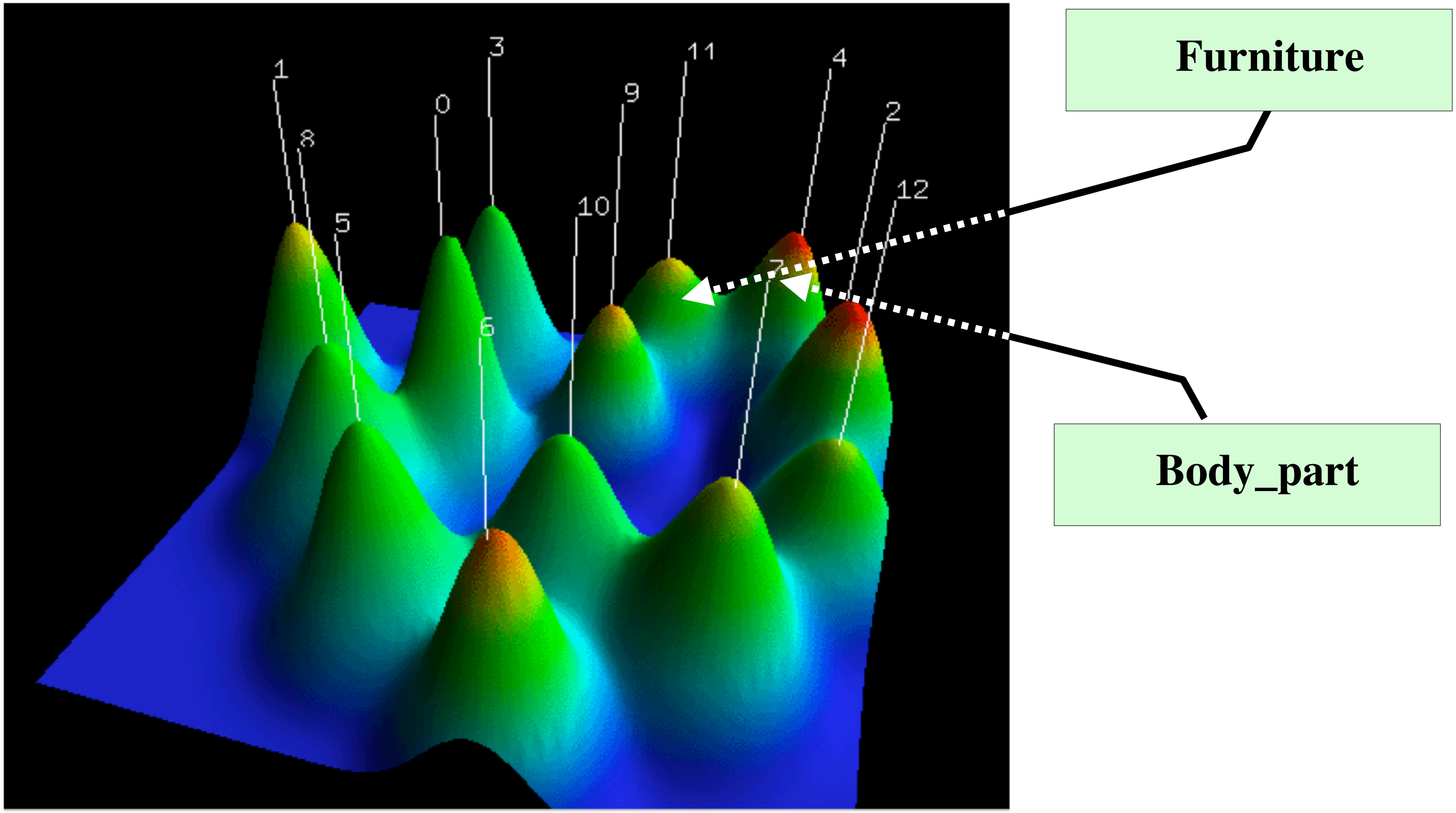
13-way clustering: [I2=9.58e+001] [214 of 214], Entropy: 0.133, Purity **0.855**

cid	Entpy	Purty	body	crea	dise	fami	vehi	publ	feel	clot	buil	time	anim	frui	furn
0	0.000	1.000	0	0	18	0	0	0	0	0	0	0	0	0	0
1	0.087	0.941	0	0	0	0	0	0	0	0	0	0	1	16	0
2	0.106	0.923	0	1	0	0	0	0	12	0	0	0	0	0	0
3	0.000	1.000	0	13	0	0	0	0	0	0	0	0	0	0	0
4	0.000	1.000	16	0	0	0	0	0	0	0	0	0	0	0	0
5	0.000	1.000	0	0	0	0	0	0	0	0	0	17	0	0	0
6	0.321	0.750	0	1	0	0	12	0	0	2	0	1	0	0	0
7	0.160	0.895	0	0	0	0	1	0	0	0	17	0	0	0	1
8	0.100	0.929	0	1	0	13	0	0	0	0	0	0	0	0	0
9	0.000	1.000	0	0	0	0	0	0	0	12	0	0	0	0	0
10	0.155	0.864	0	0	0	3	0	0	0	0	0	0	19	0	0
11	0.405	0.722	0	0	0	0	1	1	1	0	1	1	0	0	13
12	0.286	0.789	0	1	0	0	0	15	0	2	1	0	0	0	0

0.855 for Almuhareb & Poesio (2004)

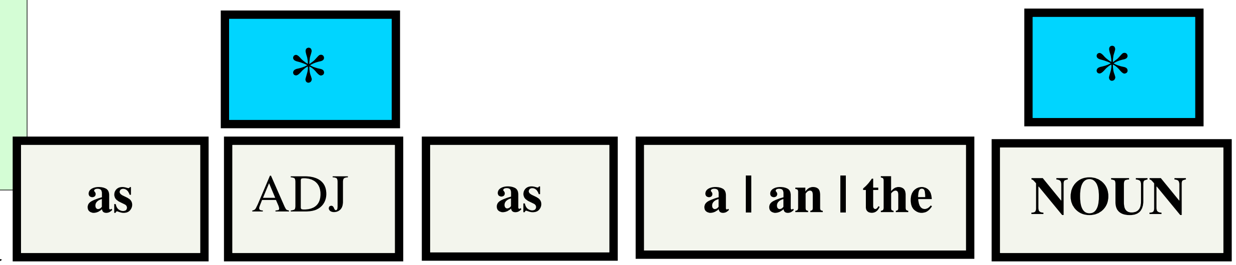
using 59,979 features

Visualizing Concept Clusters based on Diagnostic Features



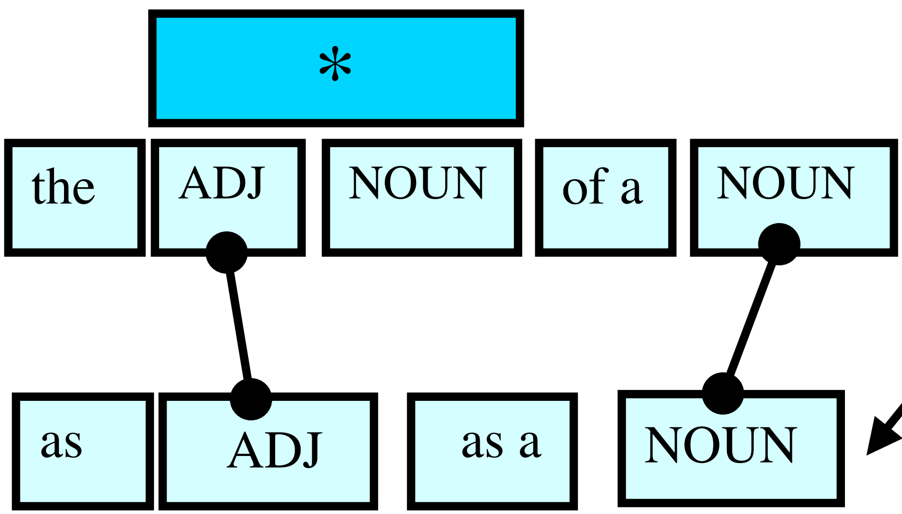
Veale & Hao (2006-08): Web-Mining of Salient Attributes from Similes

Google query to mine simile patterns for a given adj/noun



Finds 12,259 bona-fide similes, 2124 adjectives to 3778 WN noun-senses

e.g., surgeon = [delicate, skilled, precise, clinical, …]

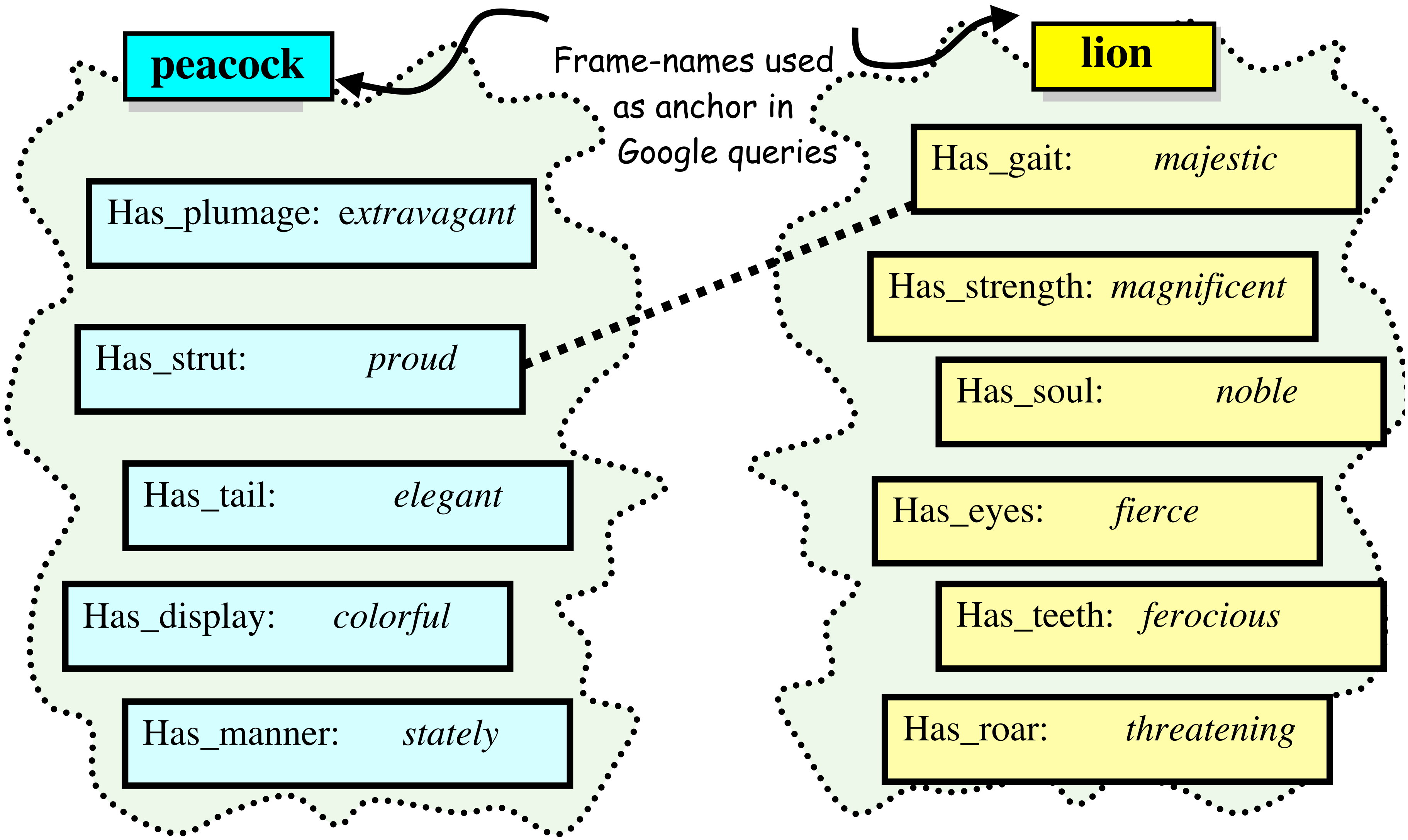


Use bindings obtained from the Simile harvest …

… to instantiate a more detailed pattern to acquire attributes for modifiers

Finds 7183 attributes for 214 nouns

Stereotypical Frames: Combining Attributes and Values



The Comparison/Simile Construction in other Languages

French	aussi (equally)	dangereux (dangerous)	qu' (as)	un (a)	requin (shark)
Spanish:	tan (as)	peligrosas (dangerous)	como (as)	un (a)	tiburón (shark)
Romanian	a fel de (equally)	periculos (dangerous)	ca si (as)		Rechin (shark)
Portuguese	tão (so)	perigoso (dangerous)	quanto (as)	um (a)	tubarão (shark)
Italian:	tanto (so much)	pericoloso (dangerous)	quanto (as)	uno (a)	squalo (shark)
Chinese:	象 (like)	鲨鱼 (shark)	一样 (equally)	危险 (dangerous)	

Veale & Hao (2007) vs. Almuhareb & Poesio (2004): Clustering Results

13-way clustering: [I2=9.58e+001] [214 of 214], Entropy: 0.133, Purity: $\boxed{0.902}$

cid	Entpy	Purty	body	crea	dise	fami	vehi	publ	feel	clot	buil	time	anim	frui	furn
0	0.000	1.000	0	0	18	0	0	0	0	0	0	0	0	0	0
1	0.087	0.941	0	0	0	0	0	0	0	0	0	0	1	16	0
2	0.106	0.923	0	1	0	0	0	0	12	0	0	0	0	0	0
3	0.000	1.000	0	13	0	0	0	0	0	0	0	0	0	0	0
4	0.000	1.000	16	0	0	0	0	0	0	0	0	0	0	0	0
5	0.000	1.000	0	0	0	0	0	0	0	0	0	17	0	0	0
6	0.321	0.750	0	1	0	0	12	0	0	2	0	1	0	0	0
7	0.160	0.895	0	0	0	0	1	0	0	0	17	0	0	0	1
8	0.100	0.929	0	1	0	13	0	0	0	0	0	0	0	0	0
9	0.000	1.000	0	0	0	0	0	0	0	12	0	0	0	0	0
10	0.155	0.864	0	0	0	3	0	0	0	0	0	0	19	0	0
11	0.405	0.722	0	0	0	0	1	1	1	0	1	1	0	0	13
12	0.286	0.789	0	1	0	0	0	15	0	2	1	0	0	0	0

Compare

0.855

for

Almuhareb

& Poesio

Compare

Simile approach

7183 features

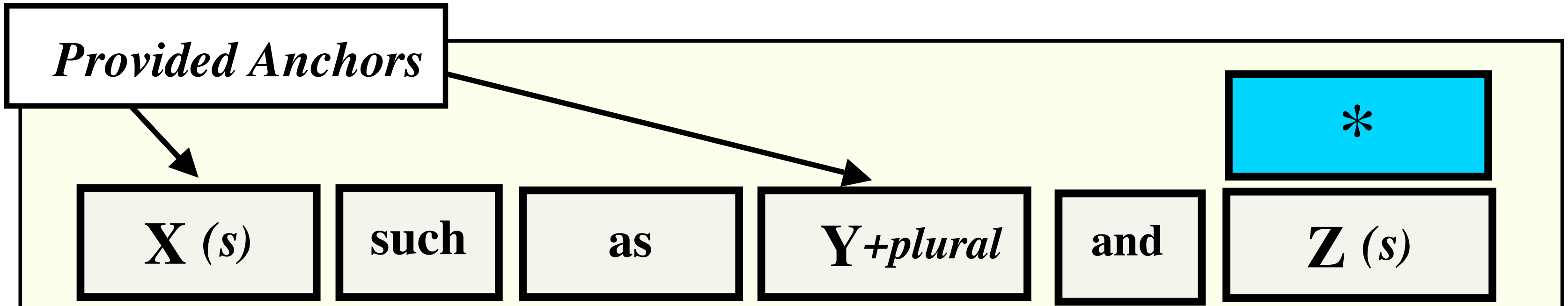
Alm.+Poesio:

59,979 features

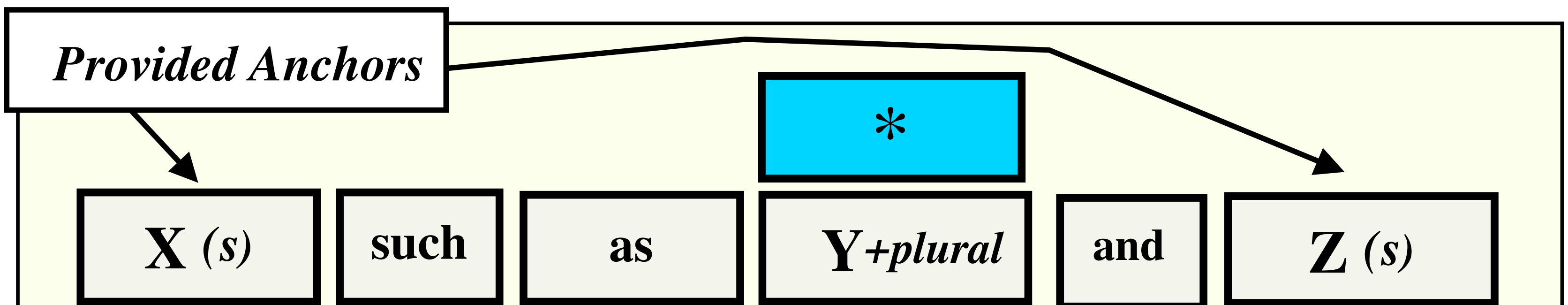
Direct Extraction Redux: Doubly-Anchored Patterns

Two "grounding" terms can be used to reduce retrieval noise

(Kozareva, Z., Riloff, E. and Hovy, E. -- ACL 2008)



E.g., "Countries such as China and Vietnam"



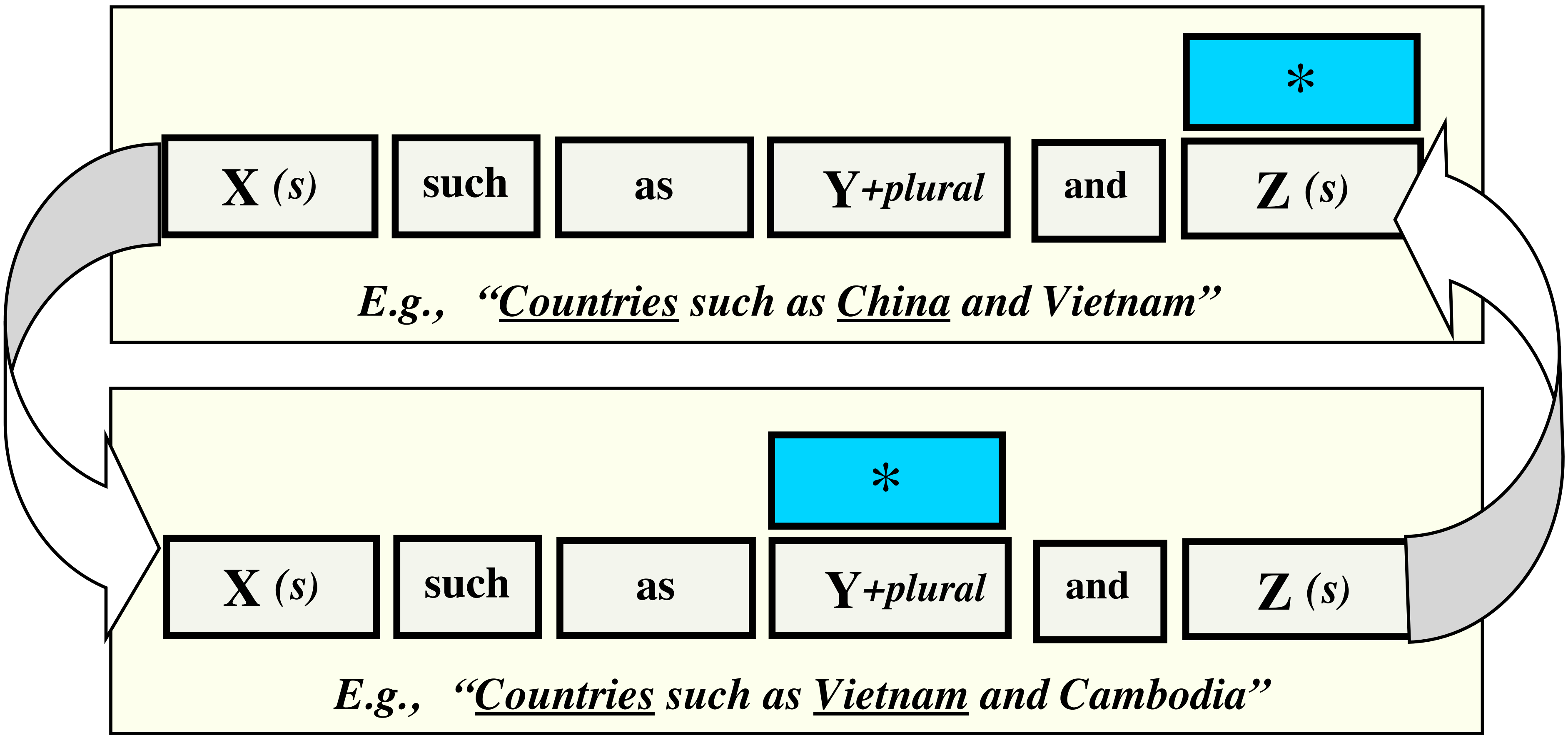
E.g., "Worms, Viruses and other malicious programs"

Useful for populating closed-classes (like Fish, Countries, etc.)

Bootstrapping with Anchored Patterns

The results of one IE cycle can be used to anchor a subsequent cycle

(Kozareva, Z., Riloff, E. and Hovy, E. -- ACL 2008)

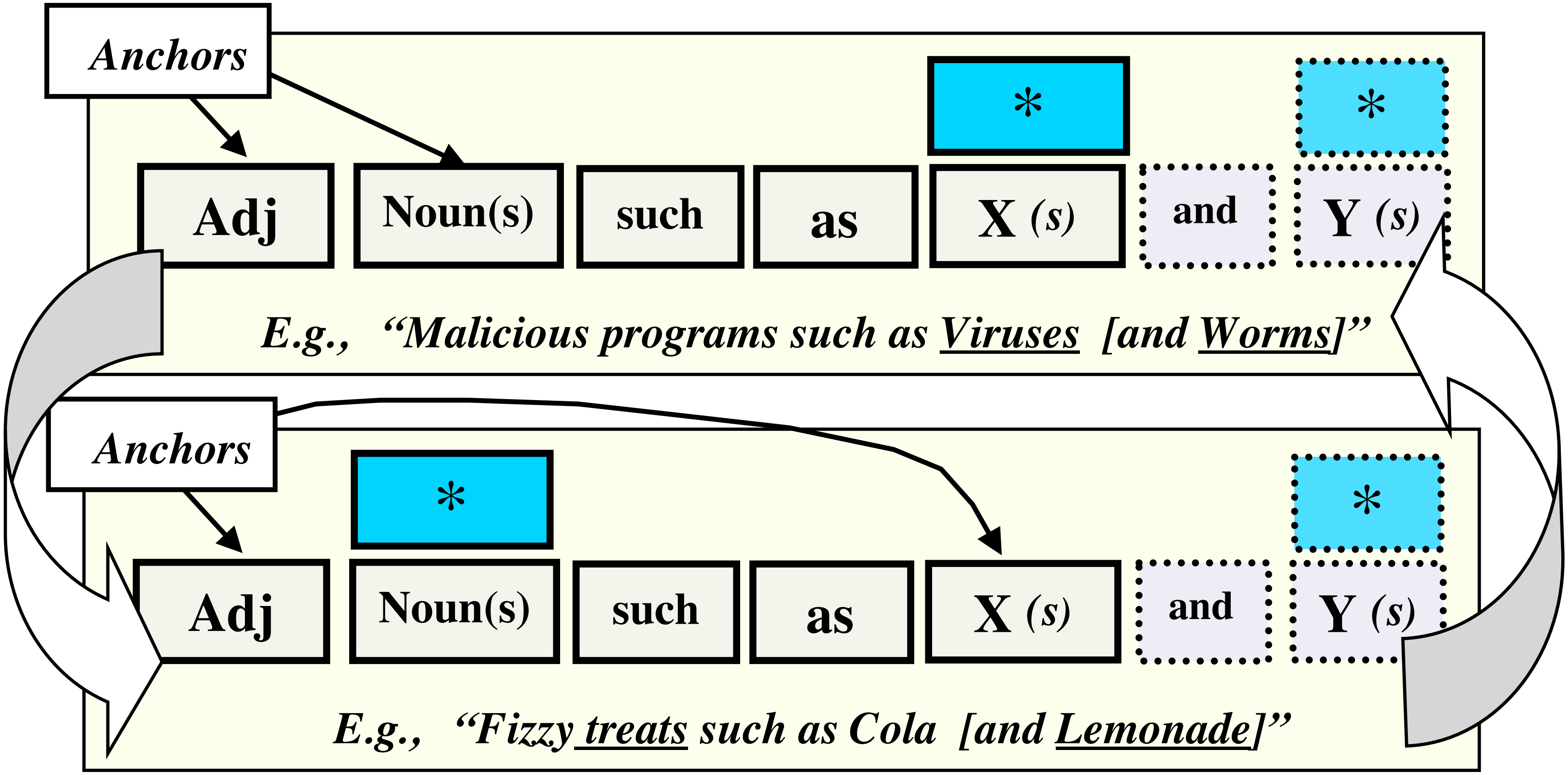


Bootstrapping can exhaustively seek out full memberships for closed sets

Bootstrapping Fine-Grained Taxonomies: Doubly-Anchored Approach

Acquiring fine-grained categories of the form Adj-Noun

e.g., triples of the form ***<cola, carbonated, drink> <cheese, soft, food>***

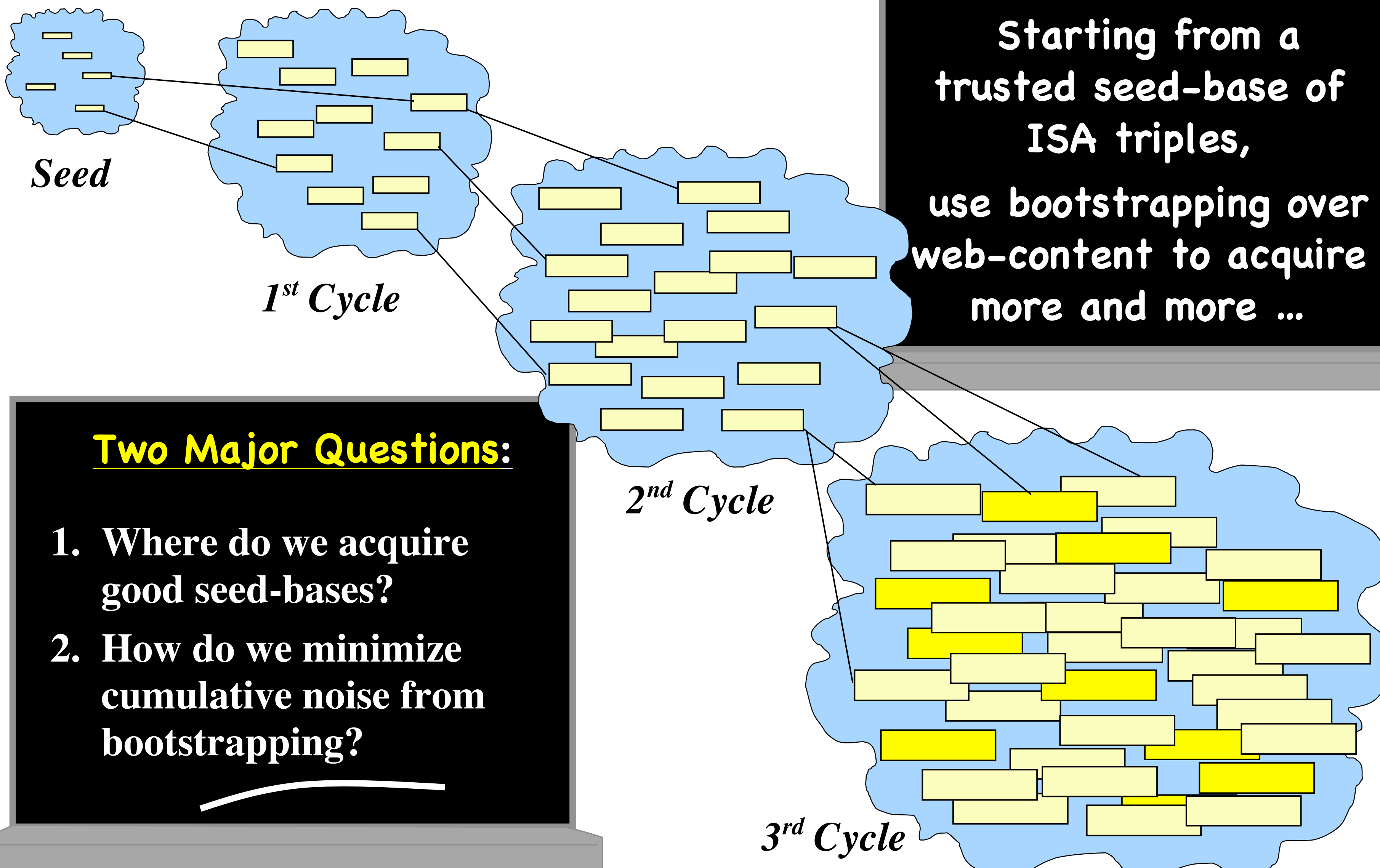


Useful for populating closed-classes (like Fish, Countries, etc.)

A Taxonomy as A Pool of Triples: How to obtain the largest Pool?

A Taxonomy as A Pool of Triples: How to obtain the largest Pool?

Starting from a trusted seed-base of ISA triples, use bootstrapping over web-content to acquire more and more …



Two Major Questions:

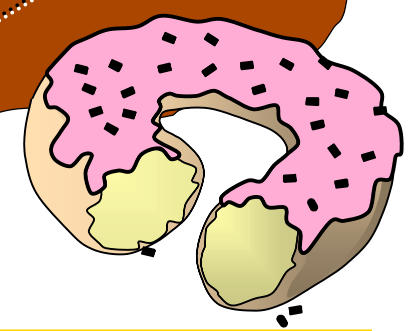
1. Where do we acquire good seed-bases?
2. How do we minimize cumulative noise from bootstrapping?

Seed # 1 (of 3) : WordNet Glosses

Shallow parse the textual glosses associated with individual WordNet senses

<espresso, black, coffee>

<espresso, strong, coffee>



E.g., **Espresso "*strong black coffee* brewed by forcing stream through …"**

Seed # 2 (of 3) : ConceptNet Propositions (IS-A)

Filter ConceptNet IS-A propositions to obtain only the most plausible ones

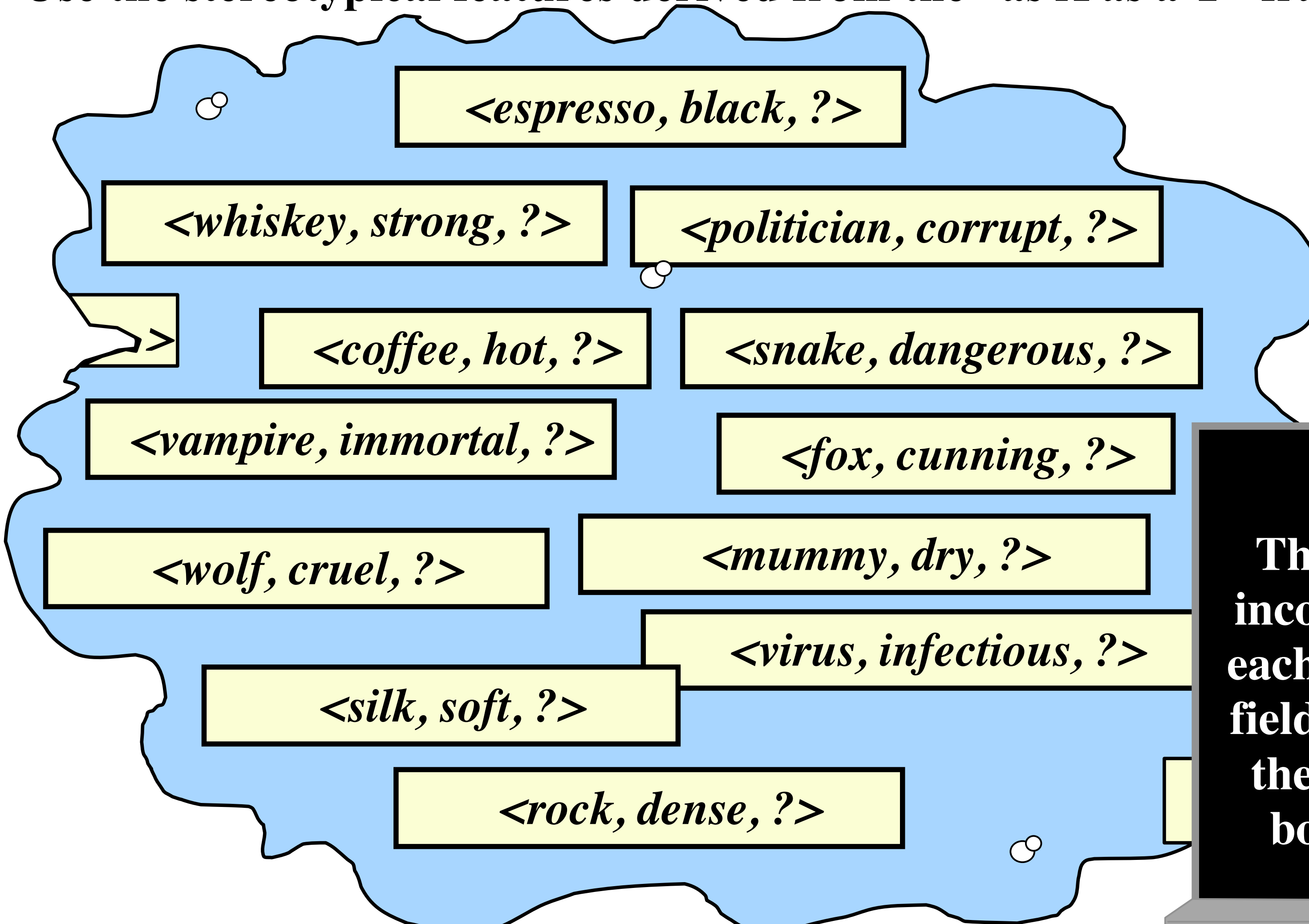
(IsA "bagdad" "capital")
(IsA "bagel" "bread")
(IsA "bagel" "breakfast food")
(IsA "bagel" "chewy kind")
(IsA "bagel" "doughnut")
(IsA "bagel" "food")
(IsA "bagel" "good food")
(IsA "bagel" "pastry")
(IsA "bagel" "roll")
(IsA "bagel" "round bread")
(IsA "bagel" "torus")
(IsA "bagpipes" "musical instrument")
(IsA "bagpipes" "scottish instrument")
(IsA "bagpipes" "traditional irish")
(IsA "bahrain" "island")

(IsA "bald eagle" "national bird")
(IsA "bald eagle" "national symbol")
(IsA "bald eagle" "national emblem")
(IsA "bald eagle" "rare bird")
(IsA "ballpoint pen" "english channel")
(IsA "bambi" "cute character")
(IsA "bambi" "ditzy name")
(IsA "bambi" "pejorative name")
(IsA "balloon" "rubber sack")
(IsA "balloon" "expensive sport")
(IsA "banana" "yellow fruit")
(IsA "banjo" "stringed instrument")
(IsA "baseball bat" "long round")
(IsA "barn" "large structure")
(IsA "baseball" "american tradition")

Find triples with *Adj-Noun* genus where Wordnet agrees with *Noun* part

Seed # 3 (of 3) : Simile-derived Associations

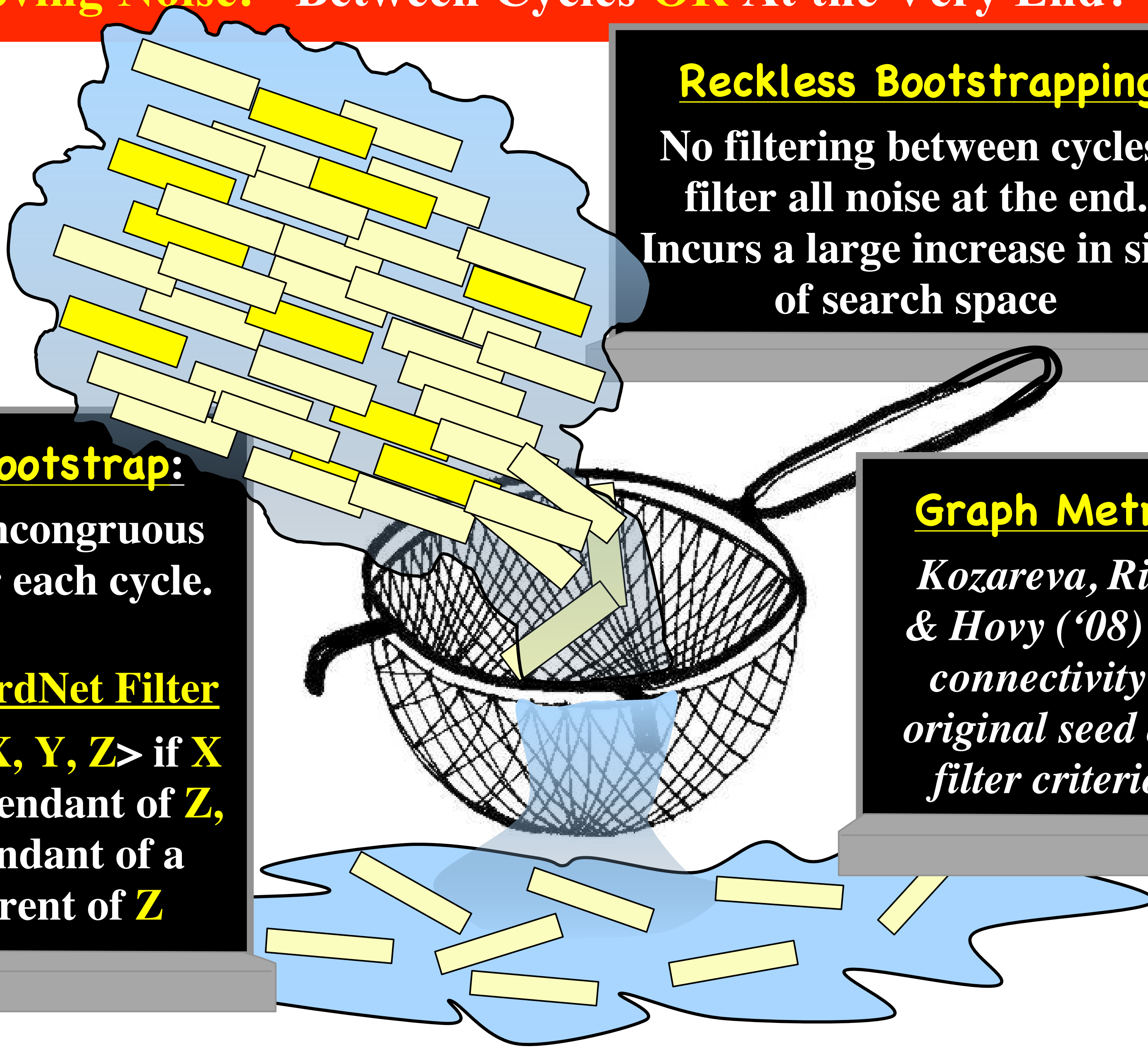
Use the stereotypical features derived from the "as X as a Y" frame earlier:



Note:

These triples are incomplete – they each **lack a parent** field, filled during the first cycle of bootstrapping

Removing Noise: Between Cycles OR At the Very End?



Reckless Bootstrapping:

No filtering between cycles – filter all noise at the end. Incurs a large increase in size of search space

Filtered Bootstrap:

Remove incongruous triples after each cycle.

Coarse WordNet Filter

Remove <X, Y, Z> if X is not a descendant of Z, or a descendant of a direct parent of Z

Graph Metrics:

Kozareva, Riloff, & Hovy ('08) use connectivity to original seed as a filter criterion

Comparing our Three Seeds: Size and Coverage

WordNet

# triples:	51,314
# terms:	12,227
# features:	2,305

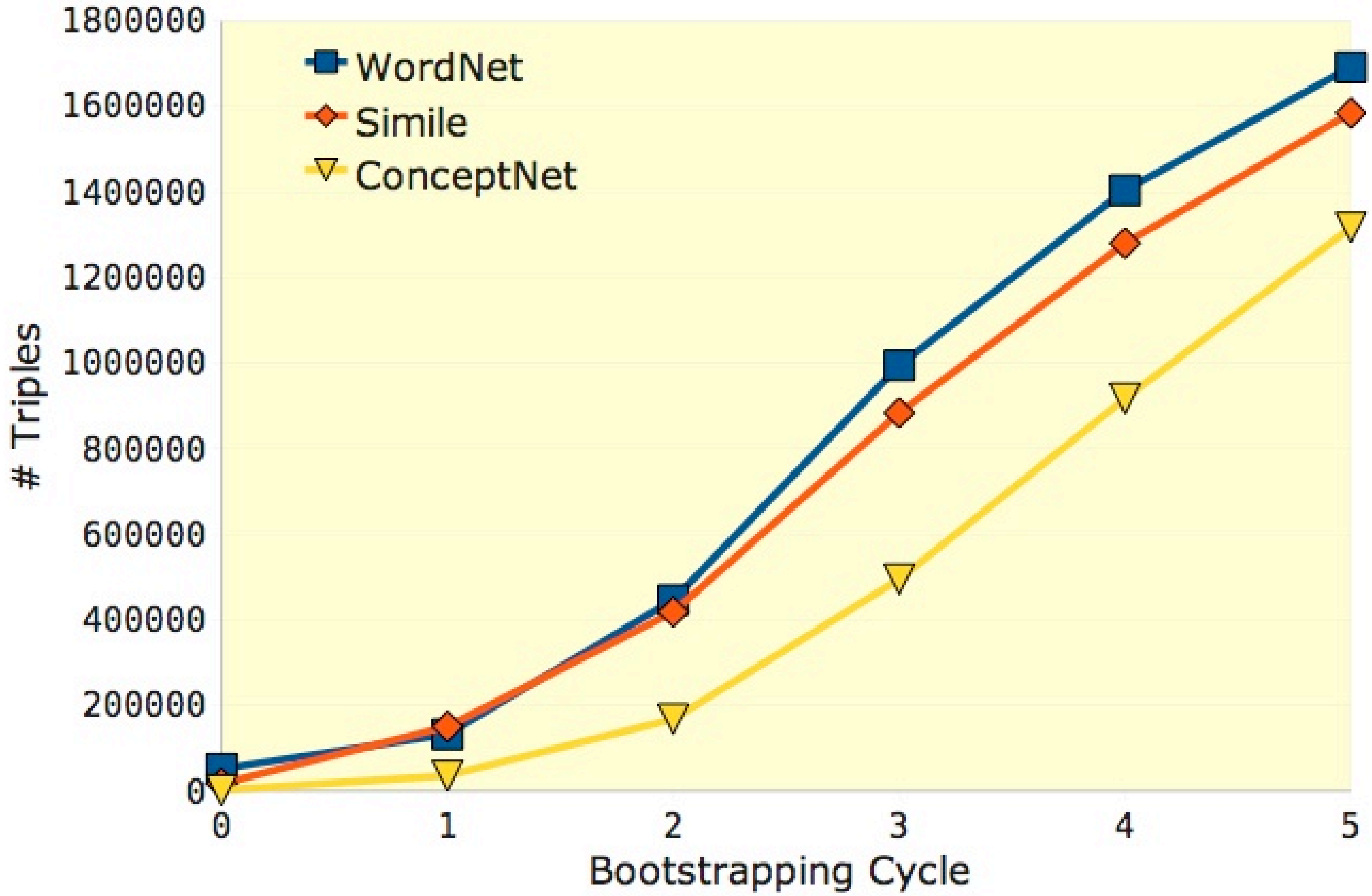
ConceptNet

# triples:	1,808
# terms:	1,133
# features:	550

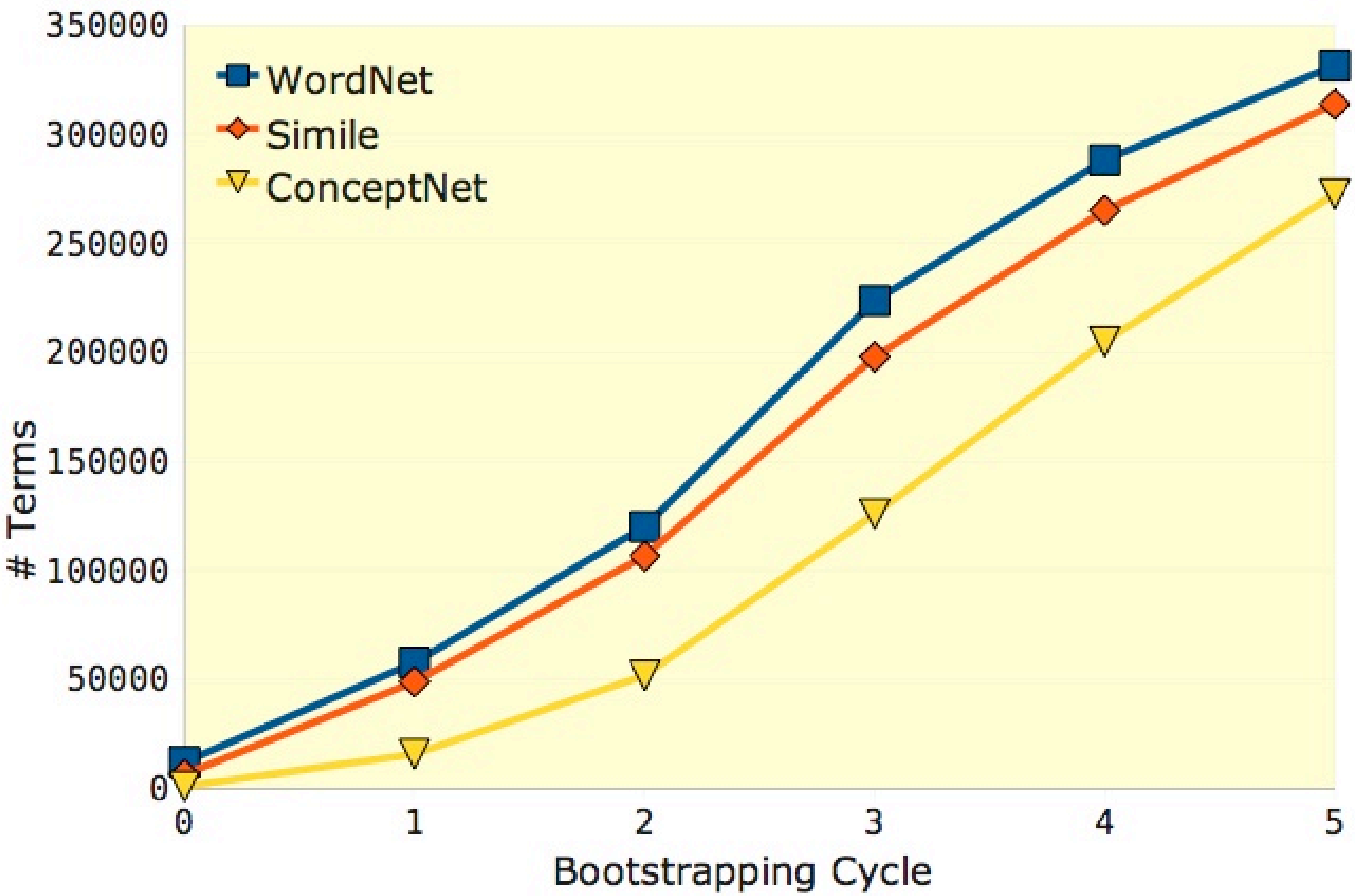
Similes

# triples:	16,688
# terms:	6,512
# features:	1,172

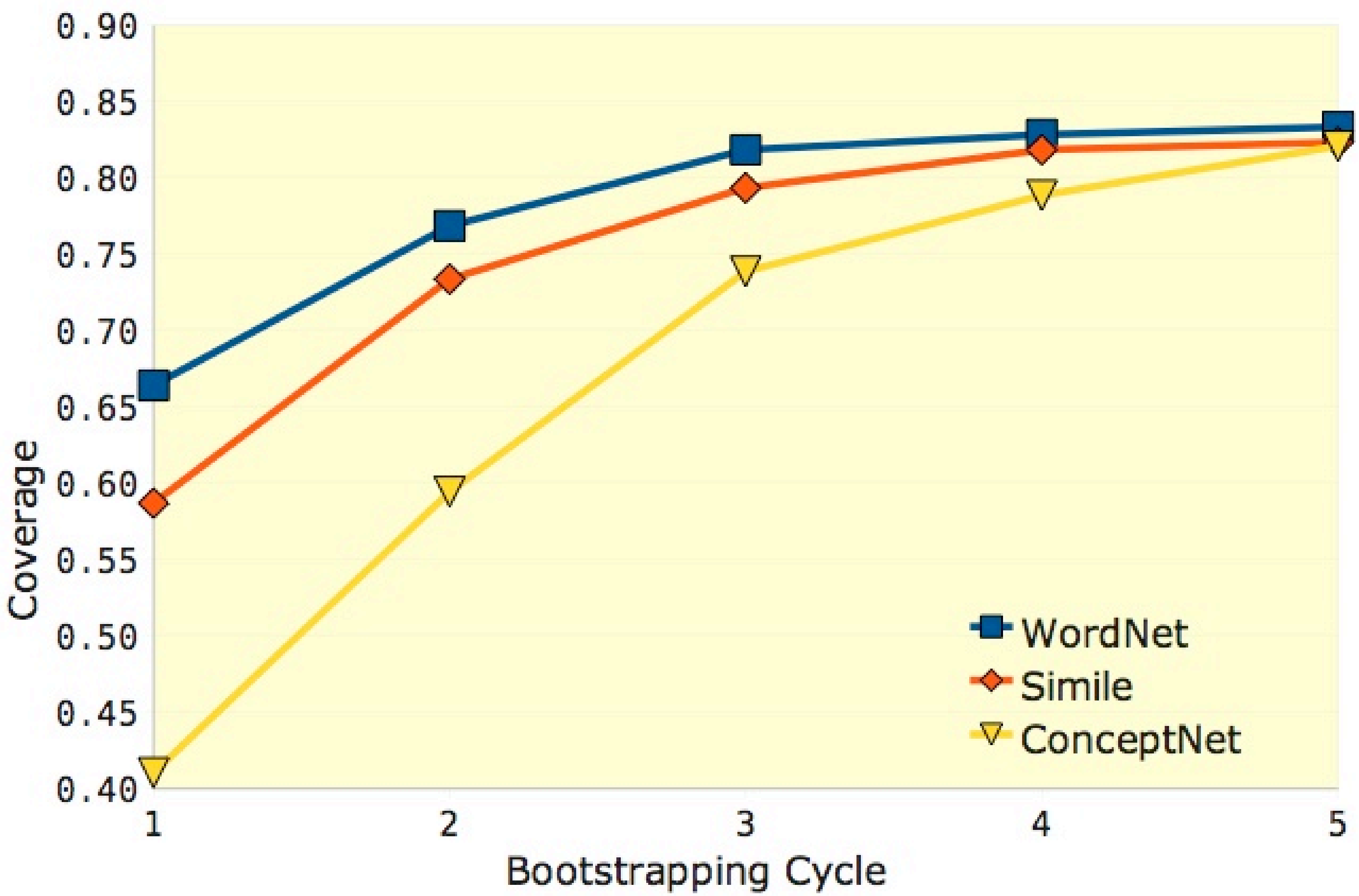
Bootstrapping Results: Growth of Structure over 5 Cycles



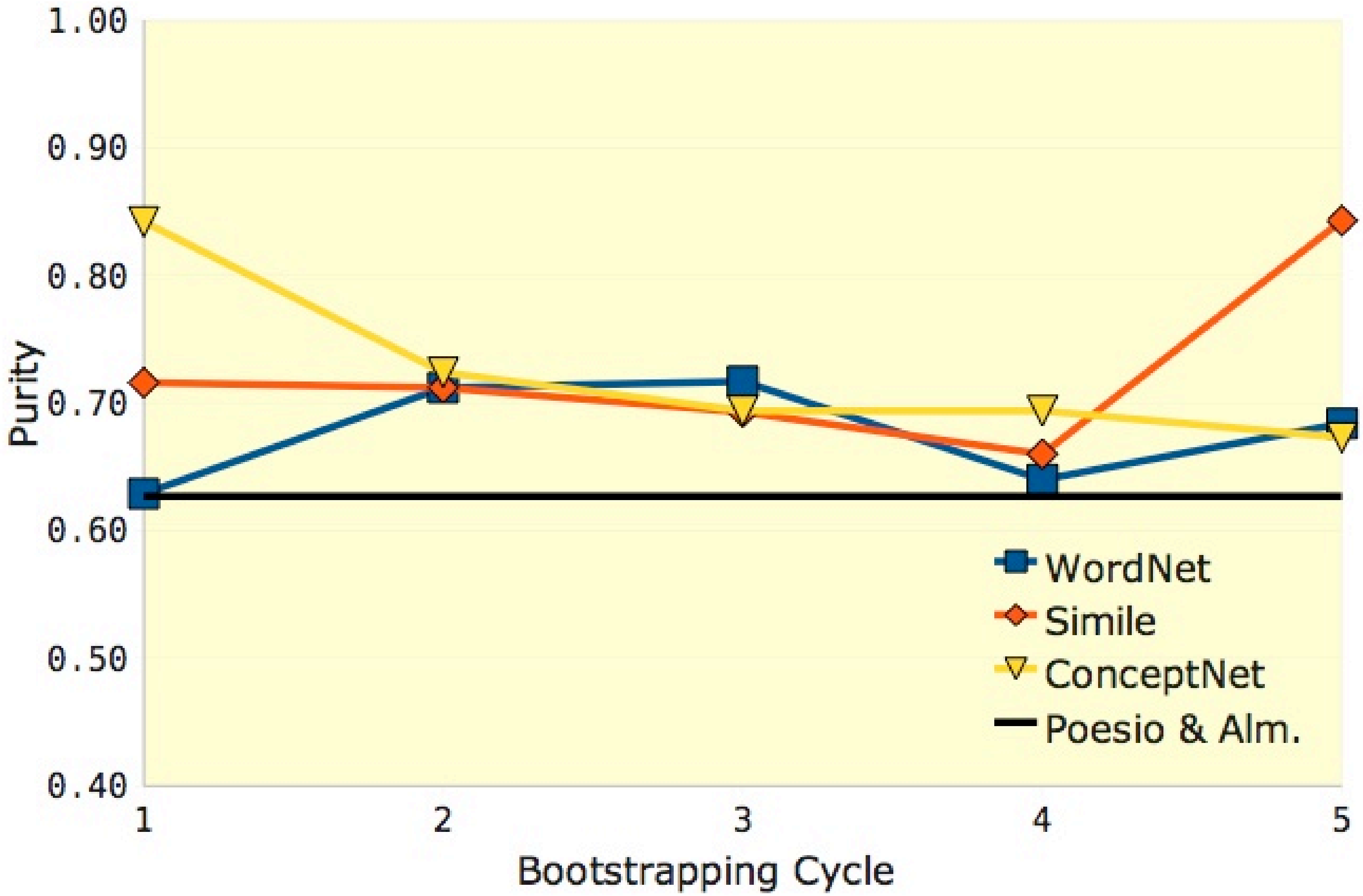
Bootstrapping Results: Accumulation of Terms over 5 Cycles



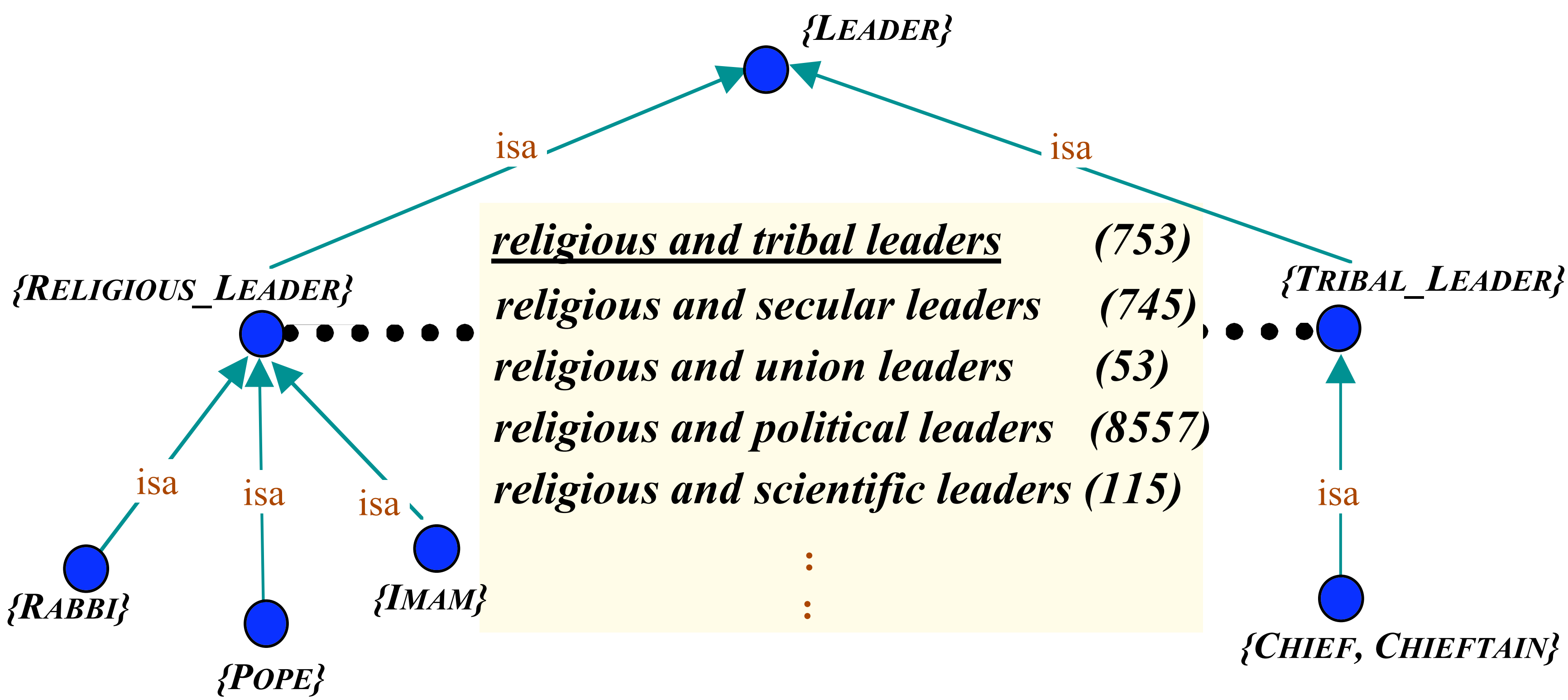
Bootstrapping Results: Increase in Coverage over 5 Cycles



Bootstrapping Results: Change in Precision over 5 Cycles

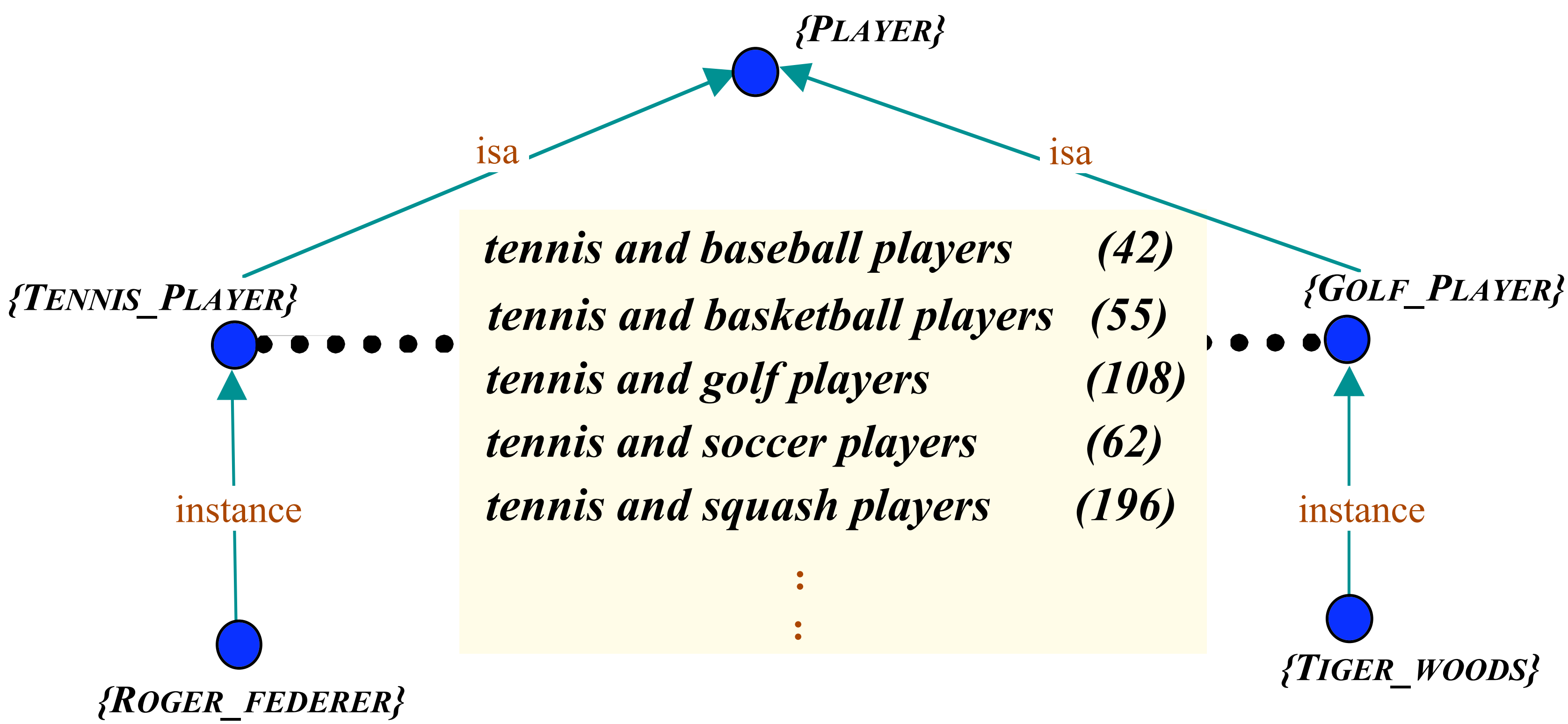


What Next: Learning Slippage Links from Corpus Data



Assume No Zeugma in compressed coordinations: find interchangeable categories

What Next: Categorizing Entities under Fine-Grained Hypernyms



Notice how modifiers cluster into semantic fields, where frequency ≈ similarity

Conclusions: Quality Wins Out over Quantity

- **Ontologies can be Constructed in a variety of different ways**

  No one approach is best: adopt an approach based on application needs

- **Handcrafted ontologies can be formally complex and knowledge-rich**

  Ironically, this richness leads to brittleness, as ontologies fail to meet goals

- **Language patterns reveal underlying ontological structure of concepts**

  Mining corpora/WWW-texts for constructions yields intuitive results

- **Large ontologies can be bootstrapped from small(-ish) seeds**

  Quality of resulting ontology depends on quality of seed, not size of seed