Never Ending Learning

Tom M. Mitchell

Justin Betteridge, Andy Carlson, Jamie Callan, William Cohen, Estevam Hruschka, Bryan Kisiel, Mahaveer Jain, Jayant Krishnamurthy, Edith Law, Thahir Mohamed, Burr Settles, Richard Wang

Machine Learning Department
Carnegie Mellon University

June 2010

Humans learn many things, for years, and become better learners over time

Why not machines?

Never Ending Learning

Task: acquire a growing competence without asymptote

- over years
- multiple functions
- where learning one thing improves ability to learn the next
- acquiring data from humans, environment

Many candidate domains:

- Robots
- Softbots
- Game players

Years of Relevant ML Research

- Architectures for problem solving/learning
  - SOAR [Newell, Laird, Rosenbloom 1986]
  - ICARUS [Langley], PRODIGY [Carbonell], …
- Life long learning, transfer learning, multi-label learning
  - EBNN [Thrun & Mitchell 1993]
  - Learning to learn [Thrun & Pratt, 1998]
- Transfer learning
  - Multitask learning [Caruana 1995]
  - Transfer reinforcement learning [Parr & Russell 1998]
  - Multilabel data [ICML 2010], Learning with structured outputs
- Active Learning
  - see survey: [Settles 2010];
  - Multi-task active learning [Harpale & Yang, ICML 2010]
- Curriculum learning
  - [Bengio, et al., 2009; Krueger & Dayan, 2009; Ni & Ling, 2010]

NELL: Never-Ending Language Learner

Inputs:

- initial ontology
- handful of examples of each predicate in ontology
- the web
- occasional interaction with human trainers

The task:

- run 24x7, forever
- each day:
    1. extract more facts from the web to populate the initial ontology
    2. learn to read (perform #1) better than yesterday

Why Do This?

1. Case study in Never-Ending Learning

2. New approach to natural language understanding
   - *Micro-reading*: sentence → content
   - *Macro-reading*: corpus, ontology → populated ontology

3. Build the world's largest structured KB
   - AI is right: intelligence requires knowledge

NELL: Never-Ending Language Learner

Goal:

- run 24x7, forever
- each day:
    1. extract more facts from the web to populate given ontology
    2. learn to read better than yesterday

Today…

Running 24 x 7, since January, 2010

Input:

- ontology defining ~200 categories and relations
- dozen seed examples of each
- 500 million web pages (ClueWeb – Jamie Callan)

Result:

- continuously growing KB with ~300,000 extracted beliefs ibm:

generalizations = {company}

candidateValues = {conference, company, product}

headquarteredIn = armonk

candidateValues = {armonk}

producesProduct = {pc}

candidateValues = {domino, thinkpad_line, ibm_e_business_logo, first_pcs, powerpc, internet, ibm_pc, iseries, rational, first_pc, quickplace, first_ibm_pc, vga_controller, original_pc, at_computer, wsfl_specification, selectric, pc, pc_convertible, workplace_client_technology, workplace, ids, opteron_server, linux_strategy, very_interesting_study, video_graphics_array, business_partner_emblem, ibm, …}

acquired = {iss, cognos, informix}

candidateValues = {spi, watchfire, telelogic, daksh, lotus, iss, internet_security_systems, gluecode, cognos, sequent, tivoli, diligent, informix, webify_solutions, geronimo, rational, information_laboratory, meiosys, webify, …}

acquiredBy = lenovo_group

candidateValues = {lenovo_group, lenovo, china, arsenal}

competesWith = {sun, texas_instruments, samsung, hewlett_packard, apple, novell, oracle, microsoft, ricoh, hp, amazon}

companyEconomicSector = {software}

hasOfficeInCountry = {united_states, canada, usa, germany, england, uk, france}

candidateValues = {san_jose, dallas, cambridge, europe, boca_raton, boulder, united_states, tucson, november, new_york, poughkeepsie, canada, october, united, research_triangle_park, rochester, beaverton, armonk, usa, u_s, germany, new_delhi, boeblingen, england, uk, france, us, facebook, masters_degree}

Semi-Supervised Bootstrap Learning

it's underconstrained!!

Extract cities:

Paris
Pittsburgh
Seattle
Cupertino

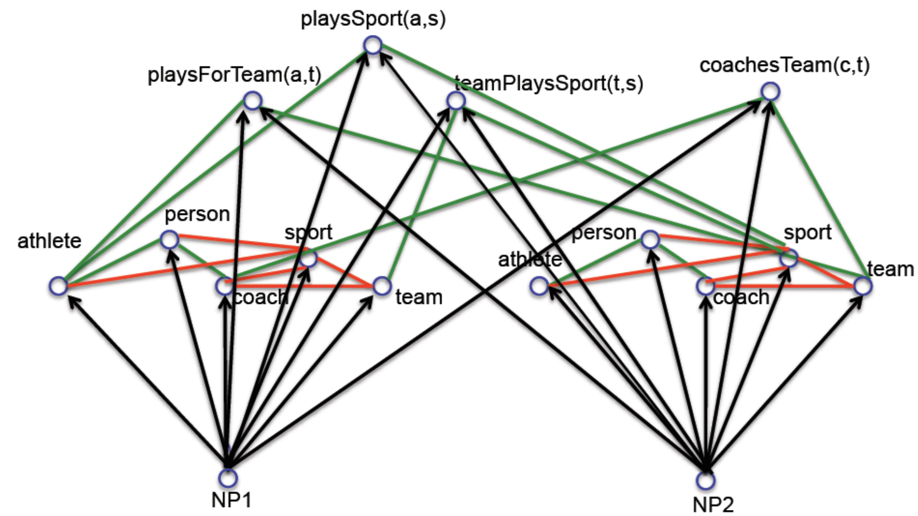
San Francisco
Austin
denial

anxiety
selfishness
Berlin

mayor of arg1
live in arg1

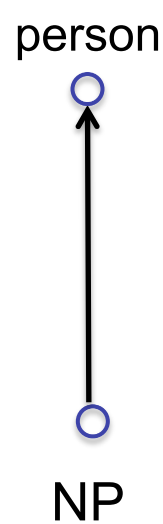
arg1 is home of
traits such as arg1

Key Idea 1: Coupled semi-supervised training of many functions



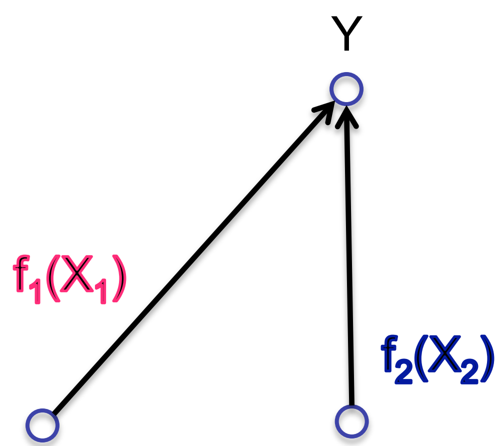
hard
(underconstrained)
semi-supervised
learning problem

much easier (more constrained)
semi-supervised learning problem


person
NP

Coupled Training Type 1: Co-Training, Multiview, Co-regularization

[Blum & Mitchell; 98]
[Dasgupta et al; 01]
[Ganchev et al., 08]
[Sridharan & Kakade, 08]
[Wang & Zhou, ICML10]

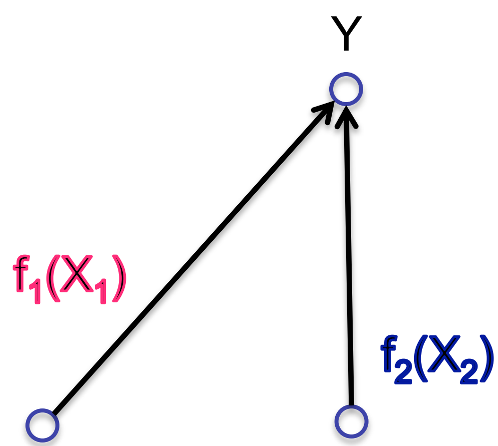


$\mathbf{X} = < X_1 \quad , \quad X_2 >$

Constraint: $f_1(x_1) = f_2(x_2)$

Coupled Training Type 1: Co-Training, Multiview, Co-regularization

[Blum & Mitchell; 98]
[Dasgupta et al; 01]
[Ganchev et al., 08]
[Sridharan & Kakade, 08]
[Wang & Zhou, ICML10]



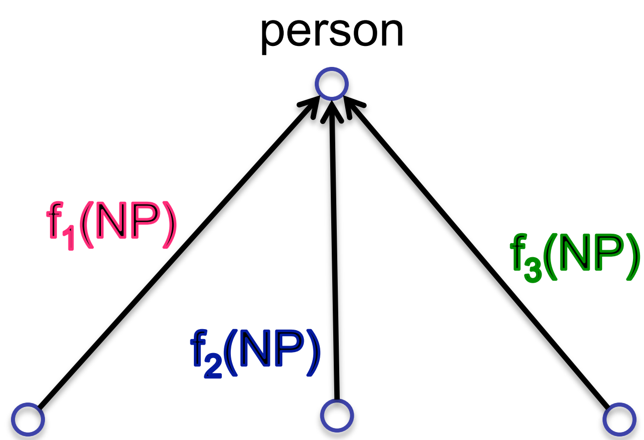
$\mathbf{X} = < X_1 , X_2 >$

Constraint: $f_1(x_1) = f_2(x_2)$

If f_1, f_2 PAC learnable,
X_1, X_2 conditionally indep
Then PAC learnable from
unlabeled data and
weak initial learner

and disagreement between
f_1, f_2 bounds error of each

Type 1 Coupling Constraints in NELL



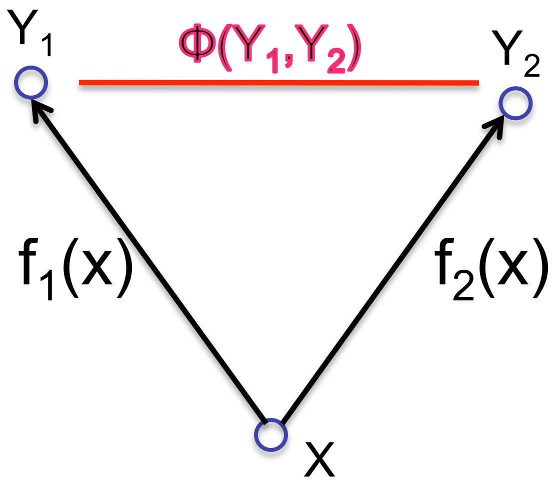
NP:

NP text context distribution	NP morphology	NP HTML contexts
__ is a friend	*capitalized?*	*www.celebrities.com:*
rang the __	*ends with '...ski'?*	* __ *
…	…	
__ walked in	*contains "univ."?*	…

Coupled training type 2
Structured Outputs, Multitask, Posterior Regularization, Multilabel

[Daume, 2008]
[Bakhir et al., eds. 2007]
[Roth et al., 2008]
[Taskar et al., 2009]
[Carlson et al., 2009]

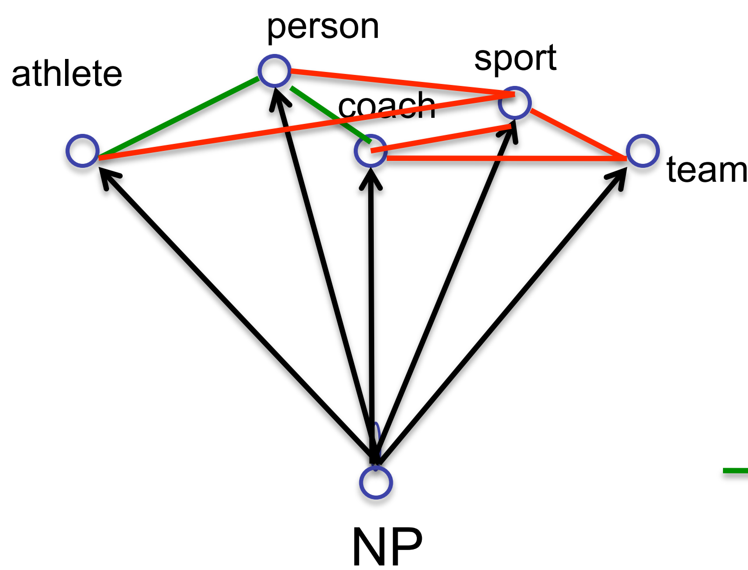
Learn functions with same input, different outputs, where we know some constraint $\Phi(Y_1,Y_2)$



Effectiveness ~ probability that $\Phi(Y_1,Y_2)$ will be violated by incorrect f_j and f_k

Constraint: $\Phi(f_1(x), f_2(x))$

Type 2 Coupling Constraints in NELL

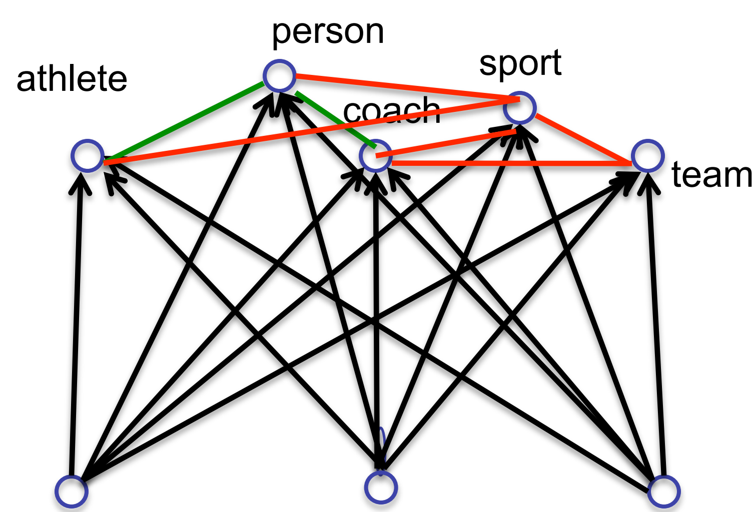


athlete(NP) → person(NP)

athlete(NP) → NOT sport(NP)

NOT athlete(NP) ← sport(NP)

Multi-view, Multi-Task Coupling

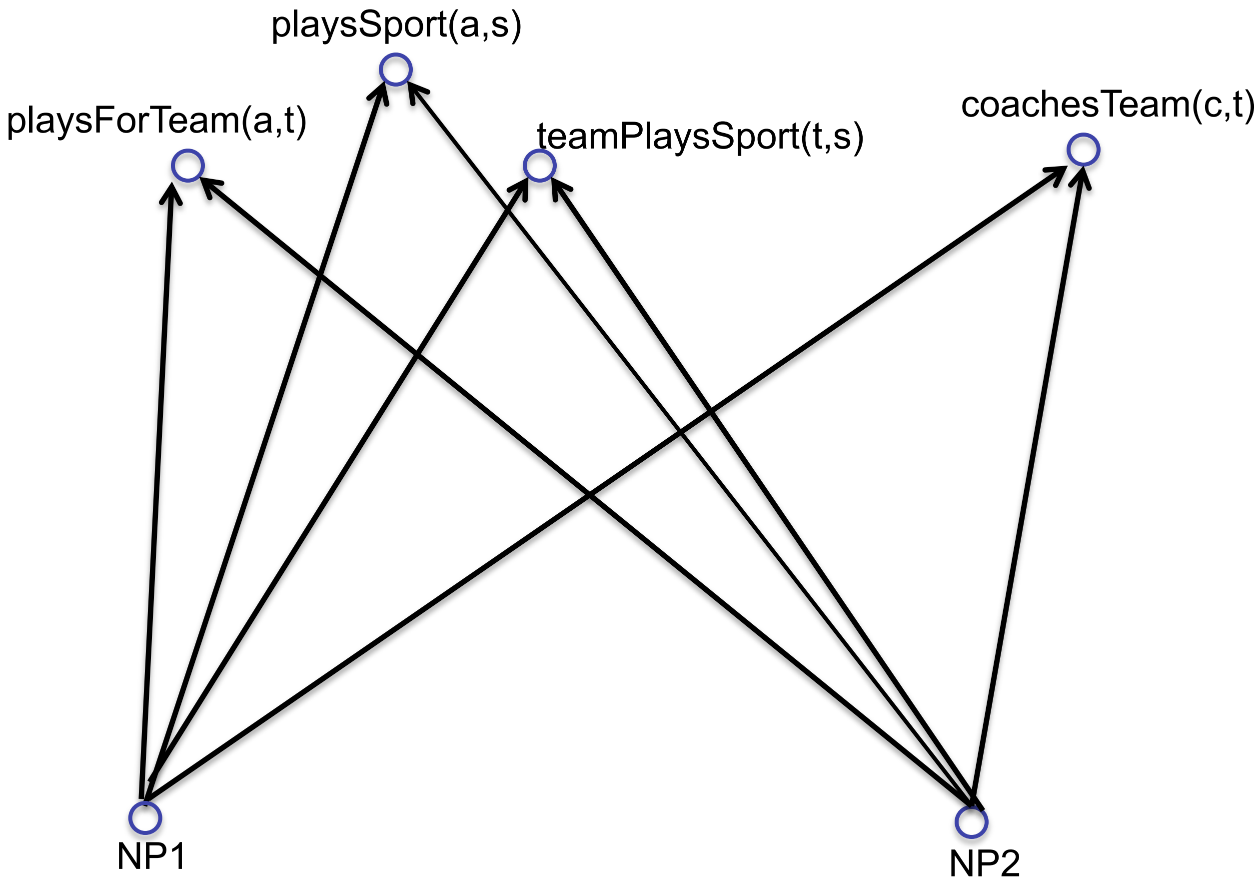


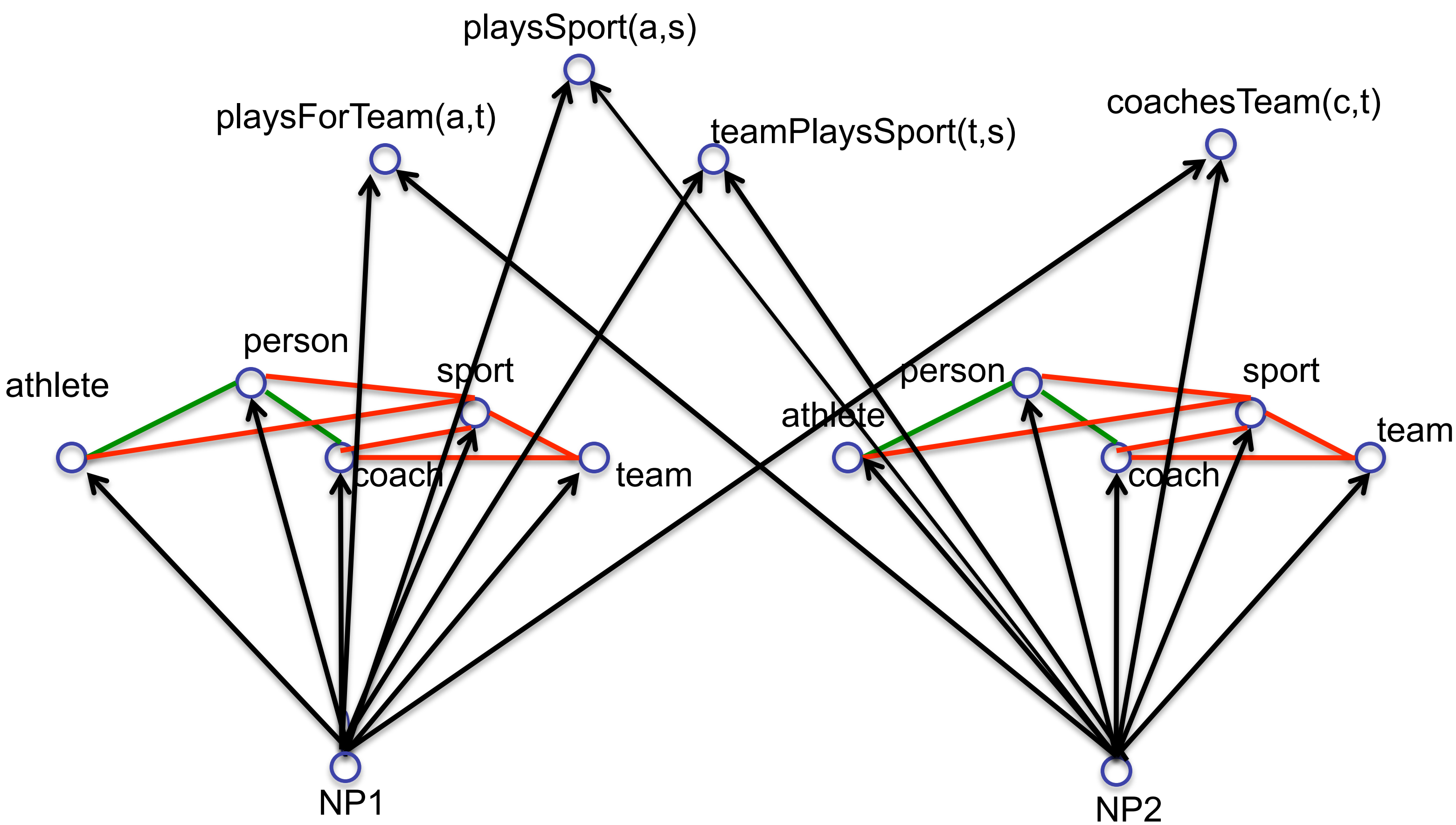
NP: NP text context distribution | NP morphology | NP HTML contexts

C categories, V views, CV = 170*3=510 coupled functions

pairwise constraints on functions ≈ 10^5

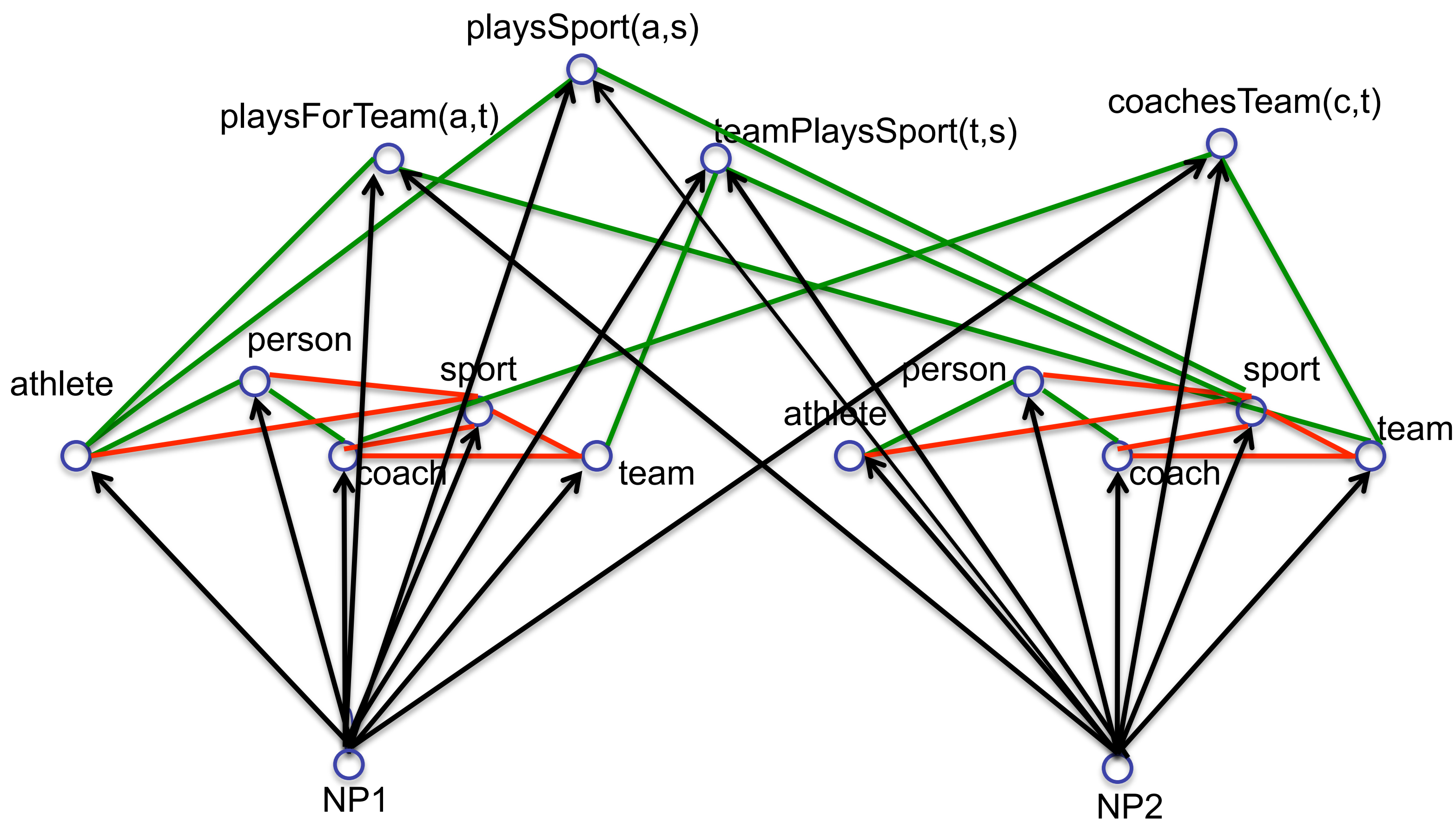
Learning Relations between NP's




playsSport(a,s)
playsForTeam(a,t)
teamPlaysSport(t,s)
coachesTeam(c,t)
person
athlete
sport
coach
team
NP1
person
athlete
sport
team
coach
NP2

Type 3 Coupling: Argument Types

Constraint: f3(x1,x2) → (f1(x1) AND f2(x2))



playsSport(NP1,NP2) → athlete(NP1), sport(NP2)

NELL: 550+ fns, coupled via 10^5 constraints

Functions

NP Morphology → fruit

NP Text Context → fruit

NP HTML Context → fruit

…

NP Morphology → city

NP Text Context → city

NP HTML Context → city

…

NP1, NP2 → mayorOf

TextContext → mayorOf

HTMLcontext → mayorOf

Constraint Vocabulary

agree(fruit(NP_Morphology), fruit(NP_TextContext))

…

mutuallyExclusive(fruit,city)

…

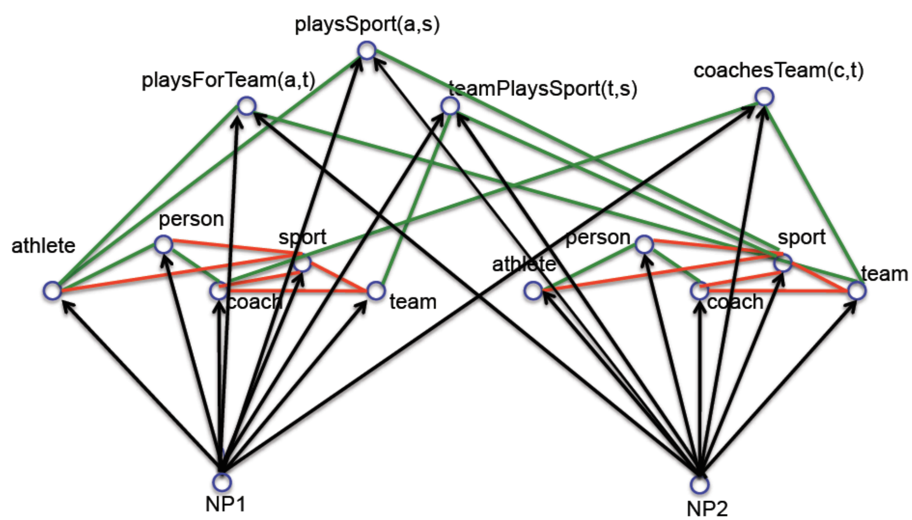
subset(city,location)

…

argumentTypes(mayorOf, city, politician)

…

Pure EM Approach to Coupled Training



E: jointly estimate latent labels for each function of each unlabeled example

M: retrain all functions, based on these probabilistic labels

Scaling problem:

- **E** step: 20M NP's, 10^{14} NP pairs to label
- **M** step: 50M text contexts to consider for each function → 10^{10} parameters to retrain
- even more URL-HTML contexts…

NELL’s Approximation to EM

E’ step:

- Consider only a growing subset of the latent variable assignments
  - category variables: up to 250 NP’s per category per iteration
  - relation variables: add only if confident and args of correct type
  - this set of explicit latent assignments <u>**IS**</u> the knowledge base

M’ step:

- Each view-based learner retrains itself from the updated KB
- “context” methods create growing subsets of contexts

NELL's Approximation to EM

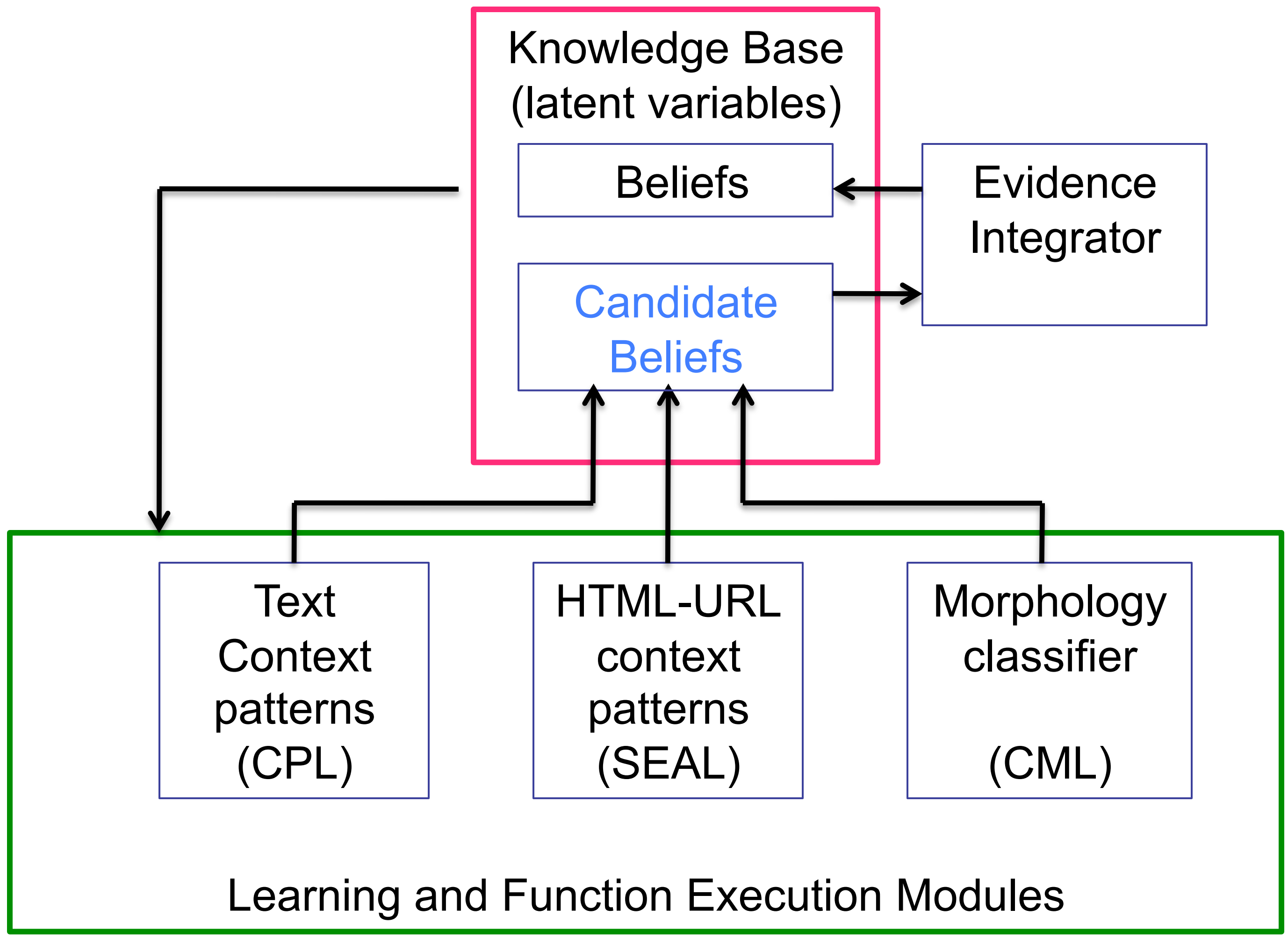
E' step:

- Consider only a growing subset of the latent variable assignments
  - category variables: up to 250 NP's per category per iteration
  - relation variables: add only if confident and args of correct type
  - this set of explicit latent assignments **IS** the knowledge base
- Assignments made in two steps
  - each view-based learner proposes candidates, probabilities
  - *Integrator* combines evidence from multiple methods and constraints, assuming independent errors

M' step:

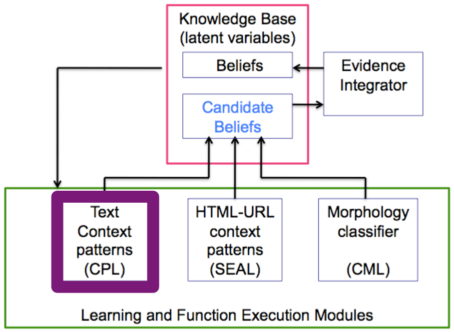
- Each view-based learner retrains itself from the updated KB
- "context" methods create growing sets of contexts

NELL Architecture



CPL

[Carlson et al., WSDM 2010]



Algorithm 1: Coupled Pattern Learner (CPL) Algorithm

Input: An ontology $\mathcal{O}$, and text corpus C
Output: Trusted instances/contextual patterns for each predicate

foreach *predicate* $p \in \mathcal{O}$ **do**
  EXTRACT new candidate instances/contextual patterns using recently promoted patterns/instances;
  FILTER candidates that violate coupling;
  RANK candidate instances/patterns;
  PROMOTE top candidates;
end

learned extraction patterns: Company

retailers_like__ such_clients_as__ an_operating_business_of__ being_acquired_by__ firms_such_as__ a_flight_attendant_for__ chains_such_as__ industry_leaders_such_as__ advertisers_like__ social_networking_sites_such_as__ a_senior_manager_at__ competitors_like__ stores_like__ __is_an_ebay_company discounters_like__ a_distribution_deal_with__ popular_sites_like__ a_company_such_as__ vendors_such_as__ rivals_such_as__ competitors_such_as__ has_been_quoted_in_the__ providers_such_as__ company_research_for__ providers_like__ giants_such_as__ a_social_network_like__ popular_websites_like__ multinationals_like__ social_networks_such_as__ the_former_ceo_of__ a_software_engineer_at__ a_store_like__ video_sites_like__ a_social_networking_site_like__ giants_like__ a_company_like__ premieres_on__ corporations_such_as__ corporations_like__ professional_profile_on__ outlets_like__ the_executives_at__ stores_such_as__ __is_the_only_carrier a_big_company_like__ social_media_sites_such_as__ __has_an_article_today manufacturers_such_as__ companies_like__ social_media_sites_like__ companies___including__ firms_like__ networking_websites_such_as__ networks_like__ carriers_like__ social_networking_websites_like__ an_executive_at__ insured_via__ __provides_dialup_access a_patent_infringement_lawsuit_against__ social_networking_sites_like__ social_network_sites_like__ carriers_such_as__ are_shipped_via__ social_sites_like__ a_licensing_deal_with__ portals_like__ vendors_like__ the_accounting_firm_of__ industry_leaders_like__ retailers_such_as__ chains_like__ prior_fiscal_years_for__ such_firms_as__ provided_free_by__ manufacturers_like__ airlines_like__ airlines_such_as__

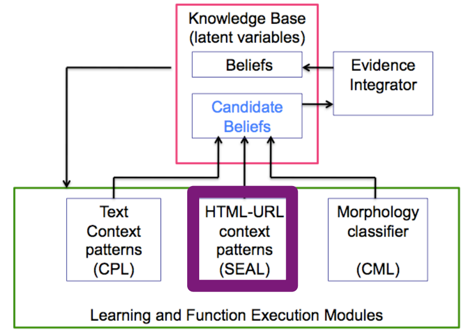
learned extraction patterns: playsSport(arg1,arg2)

arg1_was_playing_arg2 arg2_megastar_arg1 arg2_icons_arg1 arg2_player_named_arg1 arg2_prodigy_arg1 arg1_is_the_tiger_woods_of_arg2 arg2_career_of_arg1 arg2_greats_as_arg1 arg1_plays_arg2 arg2_player_is_arg1 arg2_legends_arg1 arg1_announced_his_retirement_from_arg2 arg2_operations_chief_arg1 arg2_player_like_arg1 arg2_and_golfing_personalities_including_arg1 arg2_players_like_arg1 arg2_greats_like_arg1 arg2_players_are_steffi_graf_and_arg1 arg2_great_arg1 arg2_champ_arg1 arg2_greats_such_as_arg1 arg2_professionals_such_as_arg1 arg2_course_designed_by_arg1 arg2_hit_by_arg1 arg2_course_architects_including_arg1 arg2_greats_arg1 arg2_icon_arg1 arg2_stars_like_arg1 arg2_pros_like_arg1 arg1_retires_from_arg2 arg2_phenom_arg1 arg2_lesson_from_arg1 arg2_architects_robert_trent_jones_and_arg1 arg2_sensation_arg1 arg2_architects_like_arg1 arg2_pros_arg1 arg2_stars_venus_and_arg1 arg2_legends_arnold_palmer_and_arg1 arg2_hall_of_famer_arg1 arg2_racket_in_arg1 arg2_superstar_arg1 arg2_legend_arg1 arg2_legends_such_as_arg1 arg2_players_is_arg1 arg2_pro_arg1 arg2_player_was_arg1 arg2_god_arg1 arg2_idol_arg1 arg1_was_born_to_play_arg2 arg2_star_arg1 arg2_hero_arg1 arg2_course_architect_arg1 arg2_players_are_arg1 arg1_retired_from_professional_arg2 arg2_legends_as_arg1 arg2_autographed_by_arg1 arg2_related_quotations_spoken_by_arg1 arg2_courses_were_designed_by_arg1 arg2_player_since_arg1 arg2_match_between_arg1 arg2_course_was_designed_by_arg1 arg1_has_retired_from_arg2 arg2_player_arg1 arg1_can_hit_a_arg2 arg2_legends_including_arg1 arg2_player_than_arg1 arg2_legends_like_arg1 arg2_courses_designed_by_legends_arg1 arg2_player_of_all_time_is_arg1 arg2_fan_knows_arg1 arg1_learned_to_play_arg2 arg1_is_the_best_player_in_arg2 arg2_signed_by_arg1 arg2_champion_arg1

SEAL

Set **E**xpander for **A**ny **L**anguage

[Wang and Cohen, 2007]



Seeds

ford, toyota, nissan

<li class="**ford**"><a href="http://www.curryauto.com/">

…

Extraction

<li class="honda"><a href="http://www.curryauto.com/">

honda

…

curryauto.com/">

…

<li class="**nissan**"><a href="http://www.curryauto.com/">

…

<li class="**toyota**"><a href="http://www.curryauto.com/">

…

Incorporating SEAL

[Wang and Cohen, 2007]

For each category and relation being learned,

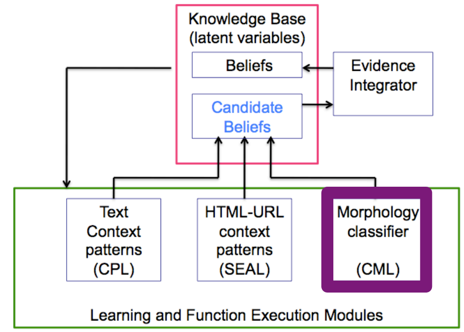
- Call Google search for sample of positive instances
- Learn URL-specific extractors for URL’s with multiple search hits

Typical learned SEAL extractors:

Predicate	Web URL	Extraction Template
academicField	http://scholendow.ais.msu.edu/student/ScholSearch.Asp	[X] -
athlete	http://www.quotes-search.com/d_occupation.aspx?o=+athlete	<a href='d_author.aspx?a=[X]'>-
bird	http://www.michaelforsberg.com/stock.html	<option>[X]</option>
bookAuthor	http://lifebehindthecurve.com/	[X] by [Y] –

CMC: Morphology Learner

[Burr Settles]



- Logistic regression classifier per predicate
- Only trained for predicates with 100 positive examples
- Negative examples from constraint propagation

Predicate	Feature	Weight
mountain	LAST=peak	1.791
mountain	LAST=mountain	1.093
mountain	FIRST=mountain	-0.875
musicArtist	LAST=band	1.853
musicArtist	POS=DT_NNS	1.412
musicArtist	POS=DT_JJ_NN	-0.807
newspaper	LAST=sun	1.330
newspaper	LAST=university	-0.318
newspaper	POS=NN_NNS	-0.798
university	LAST=college	2.076
university	PREFIX=uc	1.999
university	LAST=state	1.992
university	LAST=university	1.745
university	FIRST=college	-1.381
visualArtMovement	SUFFIX=ism	1.282
visualArtMovement	PREFIX=journ	-0.234
visualArtMovement	PREFIX=budd	-0.253

Coupled Training Helps!

[Carlson et al., WSDM 2010]

Using only two views: Text, HTML contexts.

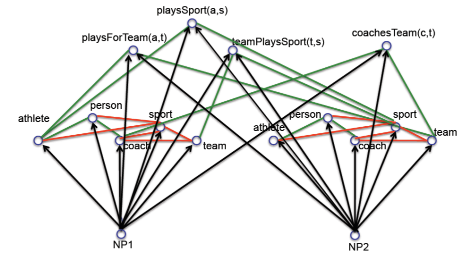
PRECISION	Text	HTML	Coupled
Categories	.41	.59	.90
Relations	.69	.91	.95

10 iterations,
200 M web pages
44 categories, 27 relations
199 extractions per category

	text	HTML	Coupled
EconomicSector	23	10	77
Emotion	53	60	83
Food	70	80	100
Furniture	0	57	90
Hobby	33	50	90
KitchenItem	3	13	100
Mammal	50	50	90
Movie	57	100	100
NewspaperCompany	60	97	100
Politician	60	37	100
Product	83	77	70
ProductType	63	63	50
Profession	53	57	93
ProfessionalOrganization	63	77	87
Reptile	3	27	100
Room	0	7	100
Scientist	30	17	100
Shape	7	7	85
Sport	13	83	73
SportsEquipment	10	23	23
SportsLeague	7	27	86
SportsTeam	30	87	87
Stadium	57	63	90
StateOrProvince	63	93	77
Tool	13	90	97
Trait	40	47	97
University	97	90	93
Vehicle	30	13	77

If coupled learning is the key idea, how can we get new coupling constraints?

Key Idea 2:



Discover New Coupling Constraints

- first order, probabilistic horn clause constraints

0.93 athletePlaysSport(?x,?y) ← athletePlaysForTeam(?x,?z) teamPlaysSport(?z,?y)

  - connects previously uncoupled relation predicates
  - infers new beliefs for KB

Discover New Coupling Constraints

For each relation:

seek probabilistic first order Horn Clauses

- Positive examples: extracted beliefs in the KB
- Negative examples: ???
  - constraints don't provide type-consistent negatives

Ontology to the rescue:

numberOfValues(teamPlaysSport) = 1

numberOfValues(competesWith) = any

can infer negative examples from positive for this, but not for this

Example Learned Horn Clauses

0.95 athletePlaysSport(?x,basketball) ← athleteInLeague(?x,NBA)

0.93 athletePlaysSport(?x,?y) ← athletePlaysForTeam(?x,?z)
teamPlaysSport(?z,?y)

0.91 teamPlaysInLeague(?x,NHL) ← teamWonTrophy(?x,Stanley_Cup)

0.90 athleteInLeague(?x,?y) ←athletePlaysForTeam(?x,?z),
teamPlaysInLeague(?z,?y)

0.88 cityInState(?x,?y) ← cityCapitalOfState(?x,?y), cityInCountry(?y,USA)

0.62* newspaperInCity(?x,New_York) ← companyEconomicSector(?x,media)
generalizations(?x,blog)

Some rejected learned rules

teamPlaysInLeague{?x nba} ← teamPlaysSport{?x basketball}

0.94 [35 0 35] [positive negative unlabeled]

cityCapitalOfState{?x ?y} ← cityLocatedInState{?x ?y}, teamPlaysInLeague{?y nba}

0.80 [16 2 23]

teamplayssport{?x, basketball} ← generalizations{?x, university}

0.61 [246 124 3063]

Rule Learning Summary

- Rule learner run every 10 iterations
- Manual filtering of rules

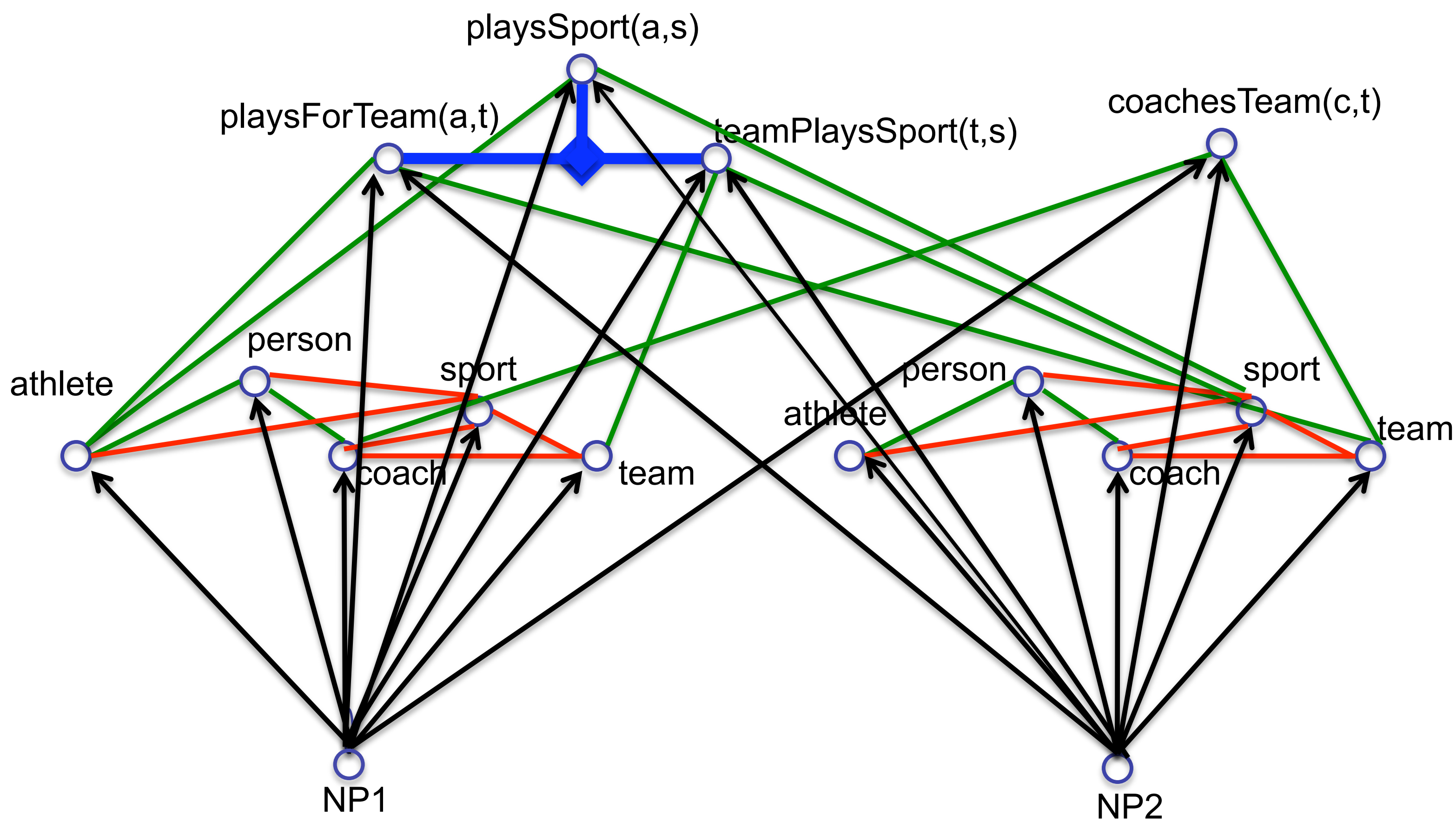
- After 120 iterations
  - 565 learned rules
  - 86% survived manual filter

  - 3948 new beliefs inferred by 486 surviving rules

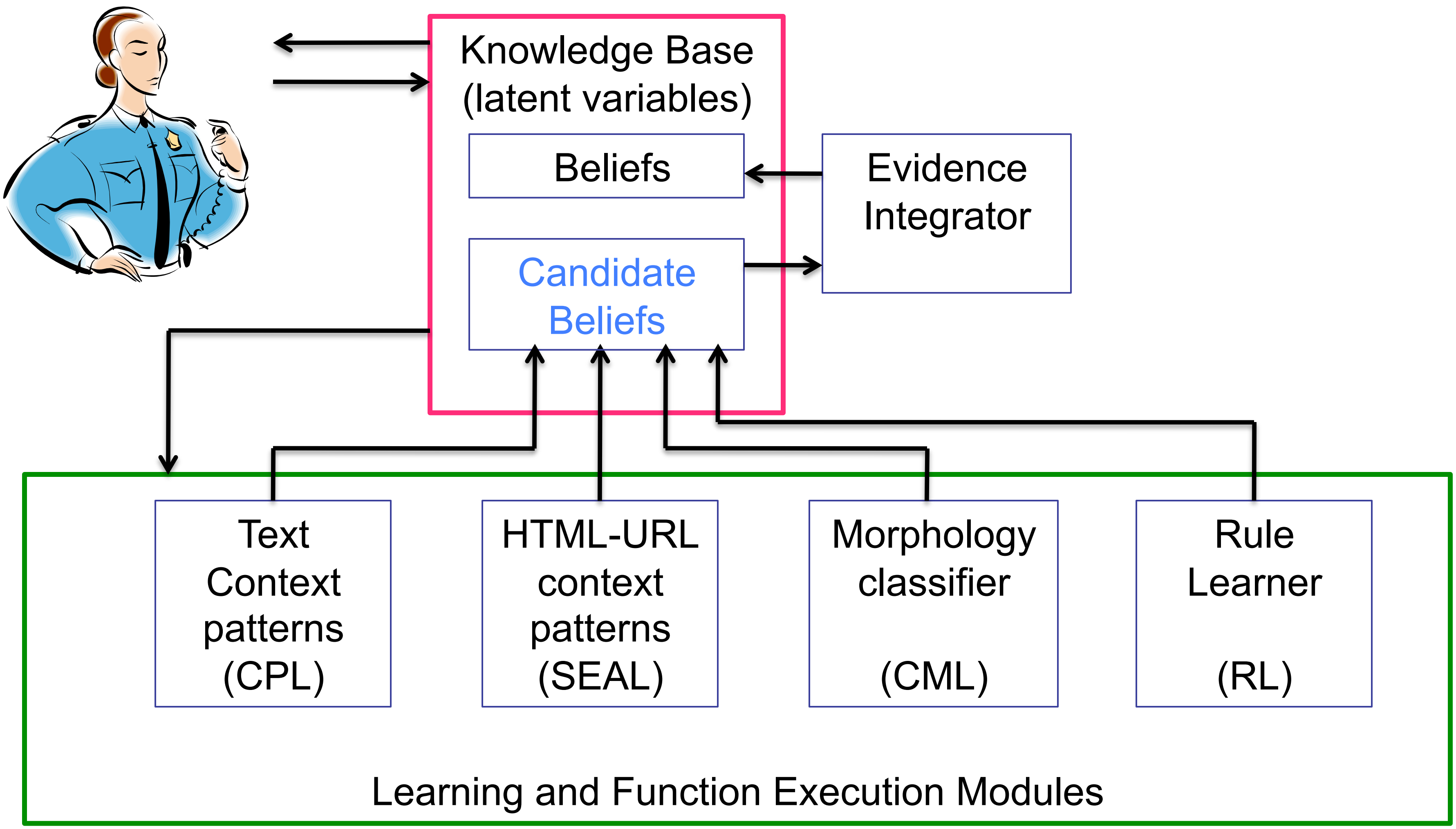
- Effectiveness limited by sparsity of relations in ontology, and restriction on numberOfValues(R)=1

Learned Probabilistic Horn Clause Rules

0.93 playsSport(?x,?y) ← playsForTeam(?x,?z), teamPlaysSport(?z,?y)

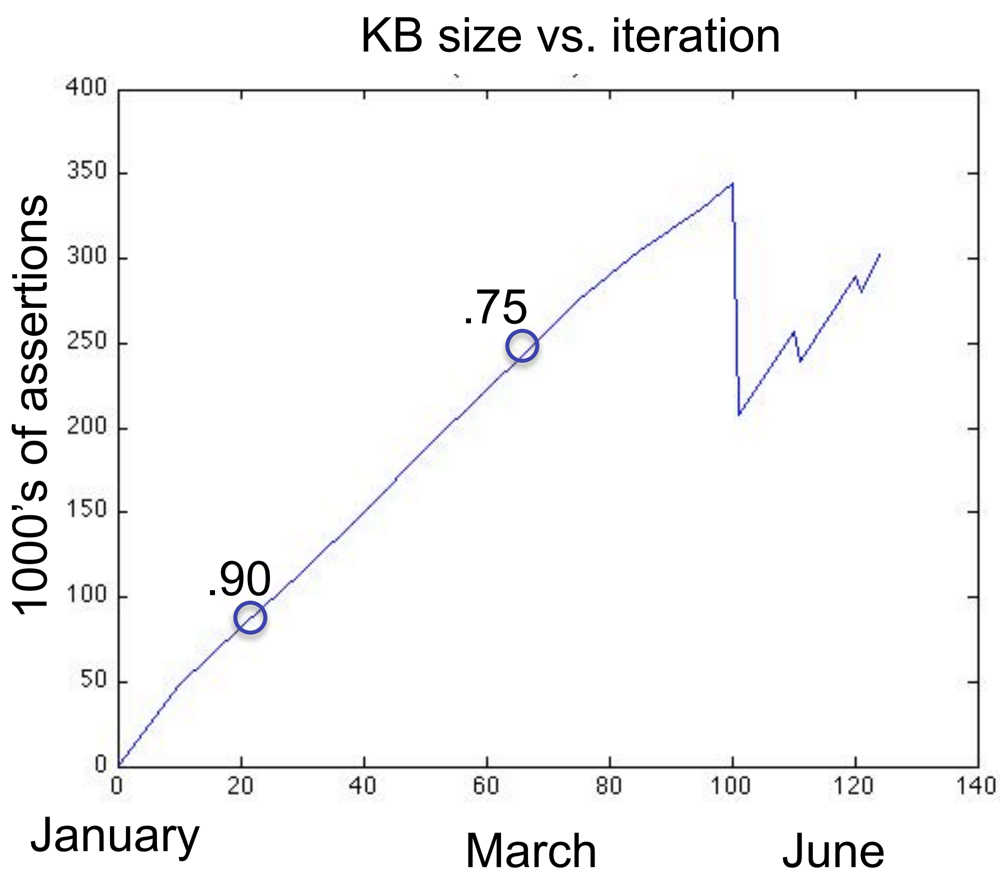


NELL Architecture, June 2010



NELL – June 2010

- 304,000 assertions
- ~30,000 learned text extraction patterns
- 486 accepted learned rules → 3948 new assertions
- Human check/clean KB every 10 iterations, beginning with iteration 100
- 65-75% of predicates currently populating well, others receiving significant correction



NELL Lessons

- Coupled semi-supervised learning of many functions helps!
- Learn new coupling constraints over time

NELL Lessons

- Coupled semi-supervised learning of many functions helps!

- Learn new coupling constraints over time

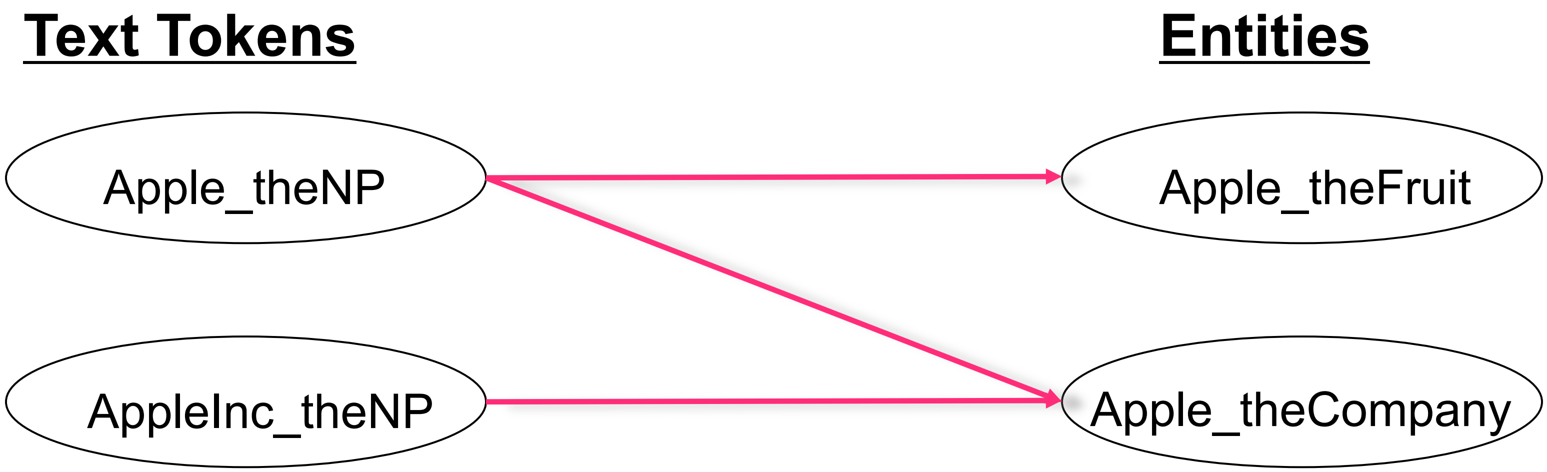
- We've changed the accuracy vs. experience learning curve from  to 

  but not to 

NELL – Next Steps

Distinguish Text Tokens from Entities

[Jayant Krishnamurthy]



Coreference Resolution:

- Co-train classifier to predict coreference as f(string similarity, extracted beliefs)
- Small amount of supervision: ~10 labeled coreference decisions
- Cluster tokens using f as similarity measure

Preliminary Coreference Results

[Jayant Krishnamurthy]

- Evaluation on “sportsteam” category
- 90% precision, 61% recall for coreference decisions

“sportsteam” Entities	Referring Tokens
St. Louis Rams	st_louis_rams, louis_rams, st___louis_rams, rams, st__louis_rams
Stanford Cardinals	stanford_university, stanford_cardinals, stanford
Pittsburgh Pirates	pittsburgh_pirates, pirates, pittsburg_pirates
Los Angeles Lakers	lakers, la_lakers, los_angeles_lakers
Valdosta State Blazers	valdosta_blazers, valdosta_st__blazers, valdosta_state_blazers
Illinois State	illinois_state, illinois_state_university, illinois_university
...	...

Ontology Extension

[Mohamed & Hruschka]

Idea:

- Discover frequently stated relations among ontology categories
- Given categories C1, C2, cluster pairs of known instances by their text contexts

* additional experiments with Etzioni & Soderland using TextRunner

Preliminary Results

[Mohamed & Hruschka]

Category Pair	Name	Text contexts	Proposed Instances
MusicInstrument Musician	Master	ARG1 master ARG2 ARG1 virtuoso ARG2 ARG1 legend ARG2 ARG2 plays ARG1	sitar , George Harrison tenor sax, Stan Getz trombone, Tommy Dorsey vibes, Lionel Hampton
Disease Disease	IsDueTo	ARG1 is due to ARG2 ARG1 is caused by ARG2	pinched nerve, herniated disk tennis elbow, tendonitis blepharospasm, dystonia
CellType Chemical	ThatRelease	ARG1 that release ARG2 ARG2 releasing ARG1	epithelial cells, surfactant neurons, serotonin mast cells, histomine
Mammals Plant	Eat	ARG1 eat ARG2 ARG2 eating ARG1	koala bears, eucalyptus sheep, grasses goats, saplings
…			

Active Learning through CrowdSourcing

[Edith Law, Burr Settles, Luis von Ahn]

mockup of *Polarity*

Two-person game to collect:

- Labels for NP's
- Information on multiple word senses and ambiguities

What will move forward research on Never Ending Learning?

Never Ending Learning: Thesis topics 1

Case study theses:

- office robot
- softbots
  - Web based research assistant
- game players
  - Why isn't there a never-ending chess learner?
- never-ending learners for sensors
  - intelligent street corner camera
  - intelligent traffic control light
  - intelligent traffic grid

Never Ending Learning: Thesis topics 2

- Scaling EM: billions of virtual(?) latent variables
  - convergence properties?
  - what properties of constraint graph predict success?
- How are correctness and self-consistency related?
  - disagreement bounds error when functions co-trained on conditionally independent features [Dasgupta, et al., 2003]
- Curriculum-based learning
  - what curriculum properties guarantee improved long term learning?
- Self-reflection:
  - what self-reflection and self-repairing capabilities assure "reachability" of target performance?

thank you!

and thanks to Yahoo! for M45 computing
and thanks to Google, NSF, DARPA for partial funding