Reading the Web: Advanced Statistical Language Processing

www.cs.cmu.edu/~tom/rtw09/

Machine Learning 10-709

September 10, 2009

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

The Plan

- What will you get out of this class?
  - knowledge of state of the art in semi-supervised learning, statistical NLP, never-ending-learning
  - an opportunity to advance it
  - a research infrastructure you might use in the future

- What will you be expected to do?
  - read, discuss, critique research papers
  - design and perform a research project in this area

- What will we build on?
  - the RTW project data and knowledge base

The Goals*

- Build the first cumulative never-ending learner

- Advance state of Natural Language Understanding

- Build and publish the world's largest structured knowledge base

* choose research problems wisely – 2/3 of success in research is (re)choosing the problem

The Problem Specification

Inputs:

- initial ontology
- handful of examples of each predicate in ontology
- the web
- occasional access to human trainer

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

What We Have Today

Goal:

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

Today…

Given:

- initial ontology defining dozens of classes and relations
- 10-20 seed examples of each

Task:

- learn to extract / extract to learn
- running over 200M web pages, for a few days

Browse the KB

- ~ 18,000+ entities, ~ 30,000 extracted beliefs
- learned from 10-20 seed examples per predicate, 200M unlabeled web pages
- ~ 2 days computation on M45 cluster

Initial ontology: Initial ontology

After a few days of self-supervised learning:

http://rtw.ml.cmu.edu/sslnlp09/index.html

http://rtw.ml.cmu.edu/wsdm10_online/

Semi-Supervised Bootstrap Learning

it's underconstrained!!

Paris
Pittsburgh
Seattle
Cupertino

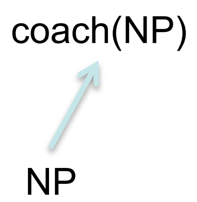
mayor of arg1
live in arg1

San Francisco
Austin
denial

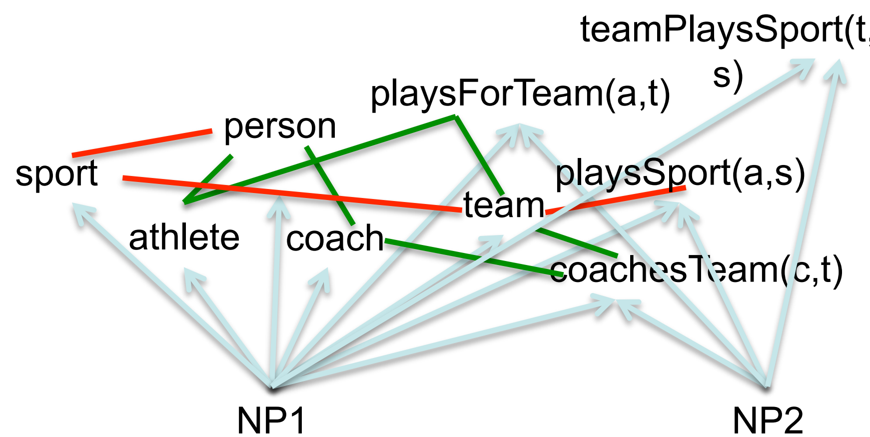
arg1 is home of
traits such as arg1

anxiety
selfishness
Berlin

The Key to Accurate Semi-Supervised Learning



Krzyzewski coaches the Blue Devils.



Krzyzewski coaches the Blue Devils.

hard (underconstrained)
semi-supervised learning
problem

much easier (more constrained)
semi-supervised learning problem

The Key: Couple the training of many functions to make unlabeled data more informative

Coupled training type 1

Wish to learn f : X → Y

e.g., city : NounPhraseInSentence → {0,1}

Coupling type 1 (co-training): Learn 2 functions with different input features

f1: X1 → Y, and f2: X2 → Y

Coupling: force their outputs to agree over unlabeled examples

city? = city?

X1 X2

X: Luke is mayor of Pittsburgh.

*co-training [Blum, Mitchell 1998]

Coupled training type 2

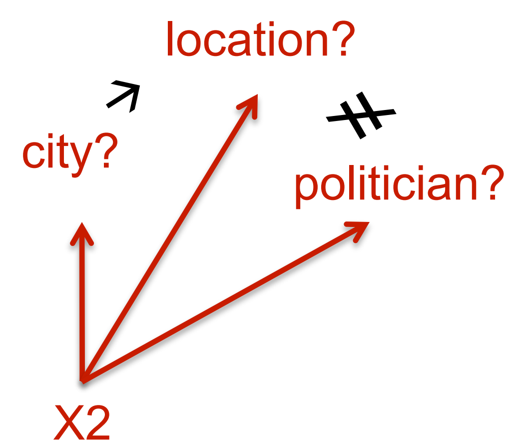
Wish to learn f: X → Y1, f: X → Y2, where g(y1,y2)

city: NounPhraseInSentence → {0,1}

politician: NounPhraseInSentence → {0,1}

Constraint type 2: force outputs to satisfy g(y1,y2)

Ontology provides coupling constraints



Luke is mayor of Pittsburgh.

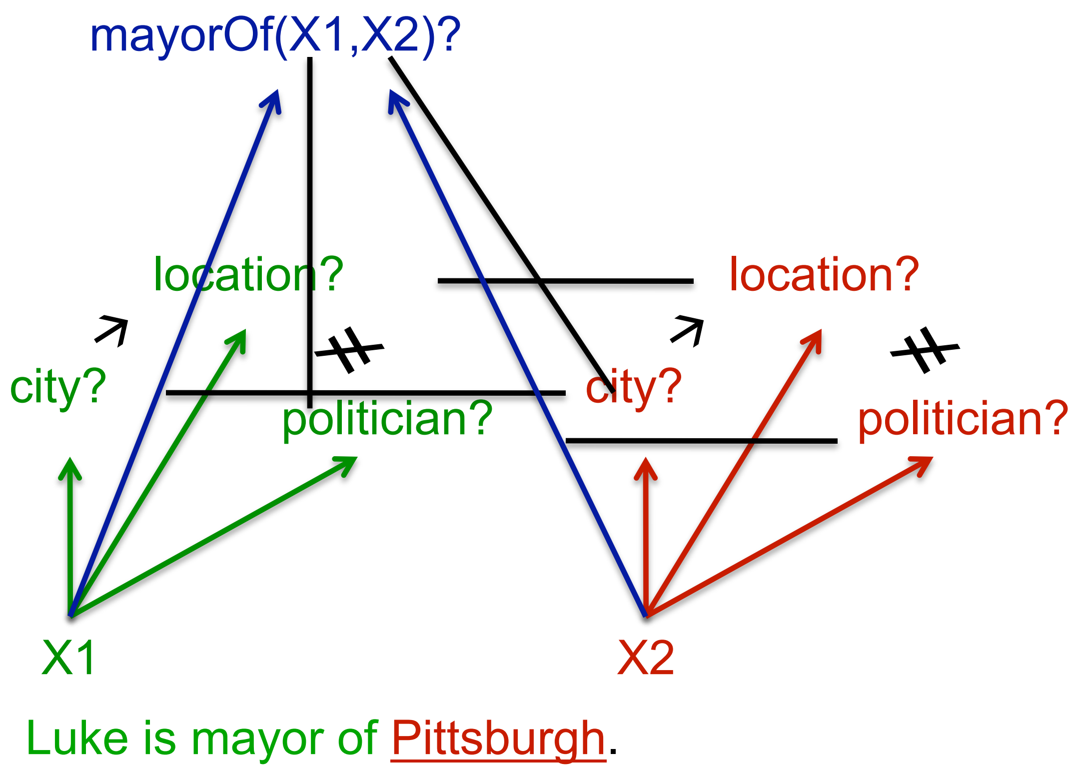
Coupled training type 3

Constraint type 3 (argument type consistency)

mayorOf: NP1 x NP2 → {0,1}

city: NP1 → {0,1}

politician: NP2 → {0,1}



Coupled Bootstrap Learner algorithm

Algorithm 1: CBL Algorithm

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

SHARE initial instances/patterns among predicates;
for $i = 1, 2, \ldots, \infty$ **do**
  foreach *predicate* $p \in \mathcal{O}$ **do**
    EXTRACT new candidate instances/patterns;
    FILTER candidates;
    TRAIN instance/pattern classifiers;
    ASSESS candidates using trained classifiers;
    PROMOTE highest-confidence candidates;
  end
  SHARE promoted items among predicates;
end

In the **ontology**: categories, relations, seed instances and patterns, type information, mutual exclusion and subset relations

Sharing enforces mutual exclusion, subset relations, and type checking

Extraction (M45):
Arg1 HQ in Arg2 → (CBC || Toronto), (Adobe || San Jose), …

Filtering (M45):
CBC || Toronto → Not enough evidence
Micron || Boise → arg2 is headquarters for chipmaker arg1,
arg1 of arg2 → too general
arg1 Corp headquarters in arg2
arg2 is headquarters for chipmaker arg1 → too specific

Assessment (M45):
Classify candidate instances with a Naive Bayes classifier
Features related to strength of co-occurrence with each pattern

Promote top-ranked instances and patterns. Use type-checking. Score patterns with estimate of precision

learned extraction patterns: Company

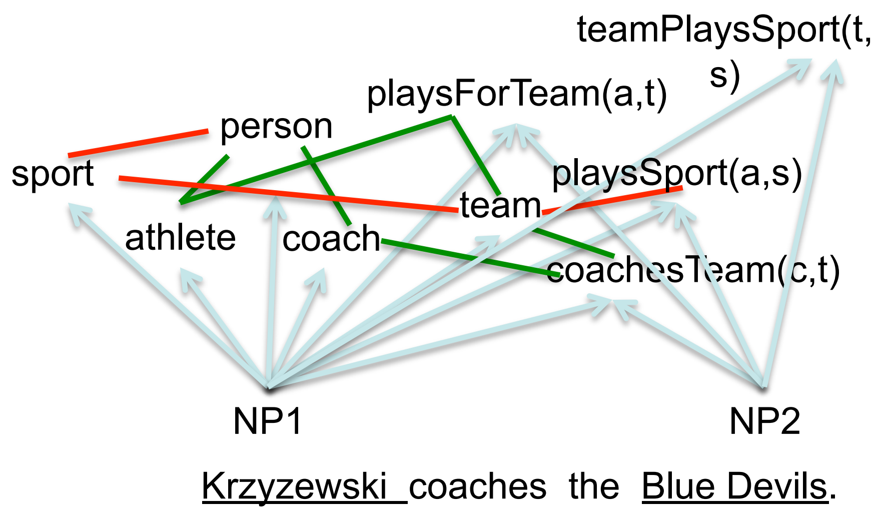
retailers_like__ such_clients_as__ an_operating_business_of__ being_acquired_by__ firms_such_as__ a_flight_attendant_for__ chains_such_as__ industry_leaders_such_as__ advertisers_like__ social_networking_sites_such_as__ a_senior_manager_at__ competitors_like__ stores_like__ __is_an_ebay_company discounters_like__ 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

If the key to accurate self-supervised learning is coupling the training of many functions,

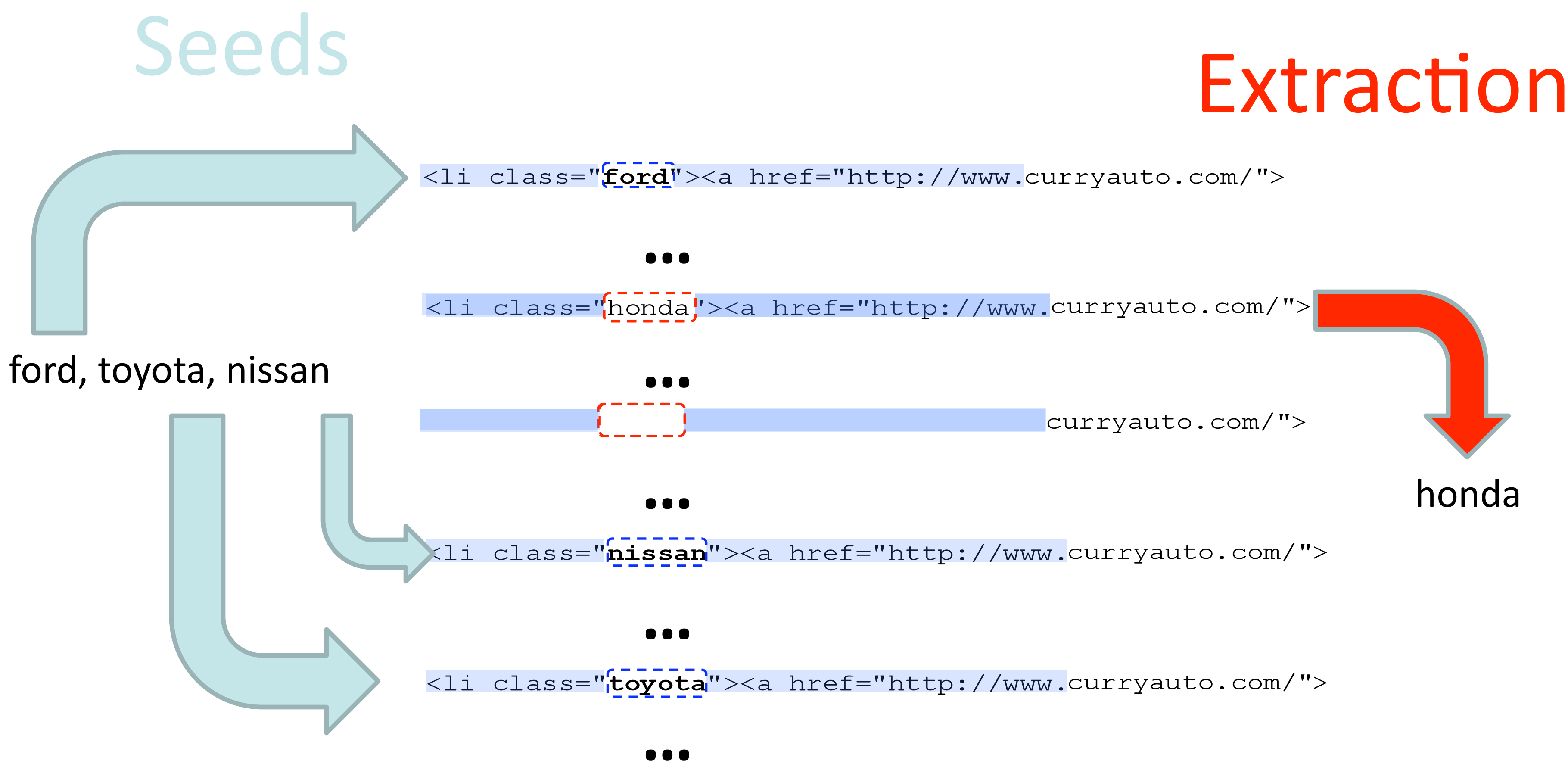
then how can we create even more coupling?



1. introduce additional coupling by adding a learner that uses HTML features instead of free text features

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

*Richard C. Wang and William W. Cohen: Language-Independent Set Expansion of Named Entities using the Web. In *Proceedings of IEEE International Conference on Data Mining* (ICDM 2007), Omaha, NE, USA. 2007.



SEAL

For each class being learned,

On each iteration

Retrain CBL from current KB, allow it to add to KB

Retrain SEAL from current KB, allow it to add to KB

Typical learned SEAL extractors:

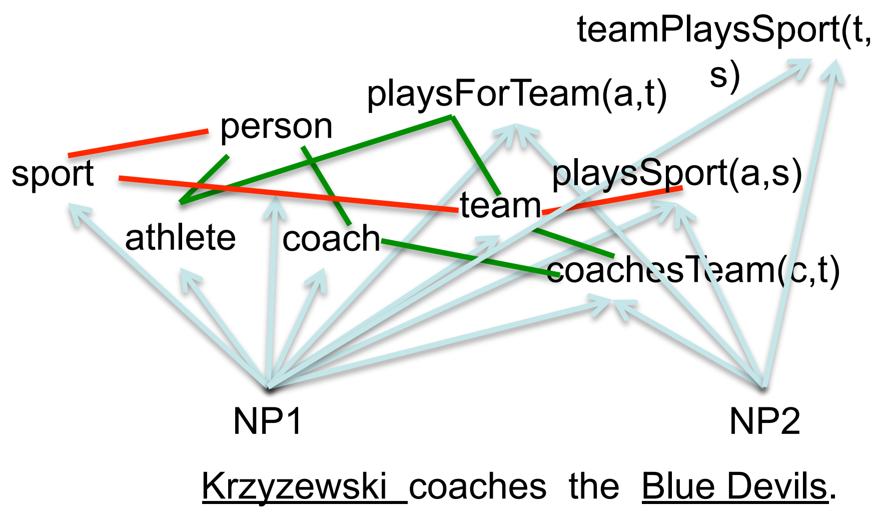
URL:	http://www.shopcarparts.com/
Wrapper:	.html" CLASS="shopcp">[...] Parts</A>
Content:	acura, audi, bmw, buick, cadillac, chevrolet, chevy, chrysler, daewoo, daihatsu, dodge, ea
URL:	http://www.allautoreviews.com/
Wrapper:	 <a href="auto_reviews/[...]/
Content:	acura, audi, bmw, buick, cadillac, chevrolet, chrysler, dodge, ford, gmc, honda, hyundai, i
URL:	http://www.hertrichs.com/
Wrapper:	<li class="franchise [...]"> <h4><a href="#">
Content:	buick, chevrolet, chrysler, dodge, ford, gmc, isuzu, jeep, lincoln, mazda, mercury, nissan,

	Precision (%)					Promoted Instances (#)				
Predicate	CPL	UPL	CSEAL	SEAL	MBL	CPL	UPL	CSEAL	SEAL	MBL
AcademicField	70	83	90	97	100	46	903	203	1000	181
Actor	100	33	100	97	100	199	1000	1000	1000	380
Animal	80	50	90	70	97	741	1000	144	974	307
Athlete	87	17	100	87	100	132	930	276	1000	555
AwardTrophyTournament	57	7	53	7	77	86	902	146	1000	79
BoardGame	80	13	70	77	90	10	907	126	1000	31
BodyPart	77	17	97	63	93	176	922	80	1000	61
Building	33	50	30	0	93	597	1000	57	1000	14
Celebrity	100	90	100	100	97	347	1000	72	747	514
CEO	33	30	100	77	100	3	902	322	1000	30
City	97	100	97	87	97	1000	1000	368	1000	603
Clothing	97	20	43	27	97	83	973	167	1000	102
Coach	93	63	100	83	100	188	838	619	1000	242
Company	97	83	100	100	97	1000	1000	245	1000	784
Conference	93	53	97	90	100	95	990	437	928	92
Country	57	33	97	37	93	1000	1000	130	1000	207
EconomicSector	60	23	100	10	77	1000	1000	34	1000	138
Emotion	77	53	87	60	83	483	992	183	1000	211
Food	90	70	97	80	100	811	1000	89	1000	272
Furniture	100	0	57	57	90	55	963	215	1000	95
Hobby	77	33	77	50	90	357	936	77	1000	127
KitchenItem	73	3	88	13	100	11	900	8	960	2
Mammal	83	50	93	50	90	224	1000	154	1000	169
Movie	97	57	97	100	100	718	1000	566	1000	183
NewspaperCompany	90	60	60	97	100	179	1000	1000	1000	241
Politician	80	60	97	37	100	178	990	30	1000	101
Product	90	83	-	77	70	1000	1000	0	999	127
ProductType	73	63	27	63	50	712	1000	31	1000	159
Profession	73	53	-	57	93	916	973	0	1000	171
ProfessionalOrganization	93	63	100	77	87	104	943	58	1000	163
Reptile	95	3	90	27	100	19	912	149	1000	54
Room	64	0	33	7	100	25	913	12	643	3
Scientist	97	30	100	17	100	83	971	928	1000	130
Shape	77	7	7	7	85	43	985	28	733	26
Sport	77	13	63	83	73	283	1000	225	1000	284
SportsEquipment	20	10	57	23	23	58	902	52	1000	174
SportsLeague	100	7	80	27	86	11	901	10	1000	14
SportsTeam	90	30	87	87	87	301	903	864	944	506
Stadium	93	57	53	63	90	102	767	944	1000	343
StateOrProvince	77	63	83	93	77	202	1000	114	1000	161
Tool	40	13	93	90	97	561	1000	713	1000	59
Trait	53	40	52	47	97	234	1000	21	1000	44
University	93	97	100	90	93	1000	1000	961	1000	516
Vehicle	67	30	50	13	77	460	1000	50	1000	98
Average	78	41	78	59	90	360	960	271	976	199
Weighted average	79	42	86	59	91					

Table 2: Precision (%) and counts of promoted instances for each category using CPL, UPL, CSEAL, SEAL MBL.

If the key to accurate self-supervised learning is <u>coupling the training of many functions</u>,

then how can we create even more coupling?



2. allow learner to discover new coupling constraints

(by mining its extracted beliefs)

Learned Probabilistic Horn Clause Rules

- 40 learned rules for teamPlaySport, playSport,
- when applied, inferred over 800 new beliefs
  - e.g., teamPlaysSport(Caps,hockey),
  - playSport(JasonGiambi,baseball)

0.84 playsSport(?x,?y) ← playsFor(?x,?z), teamPlaysSport(?z,?y)

0.70 playsSport(?x,baseball) ← playsFor(?x,cubs)

…

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

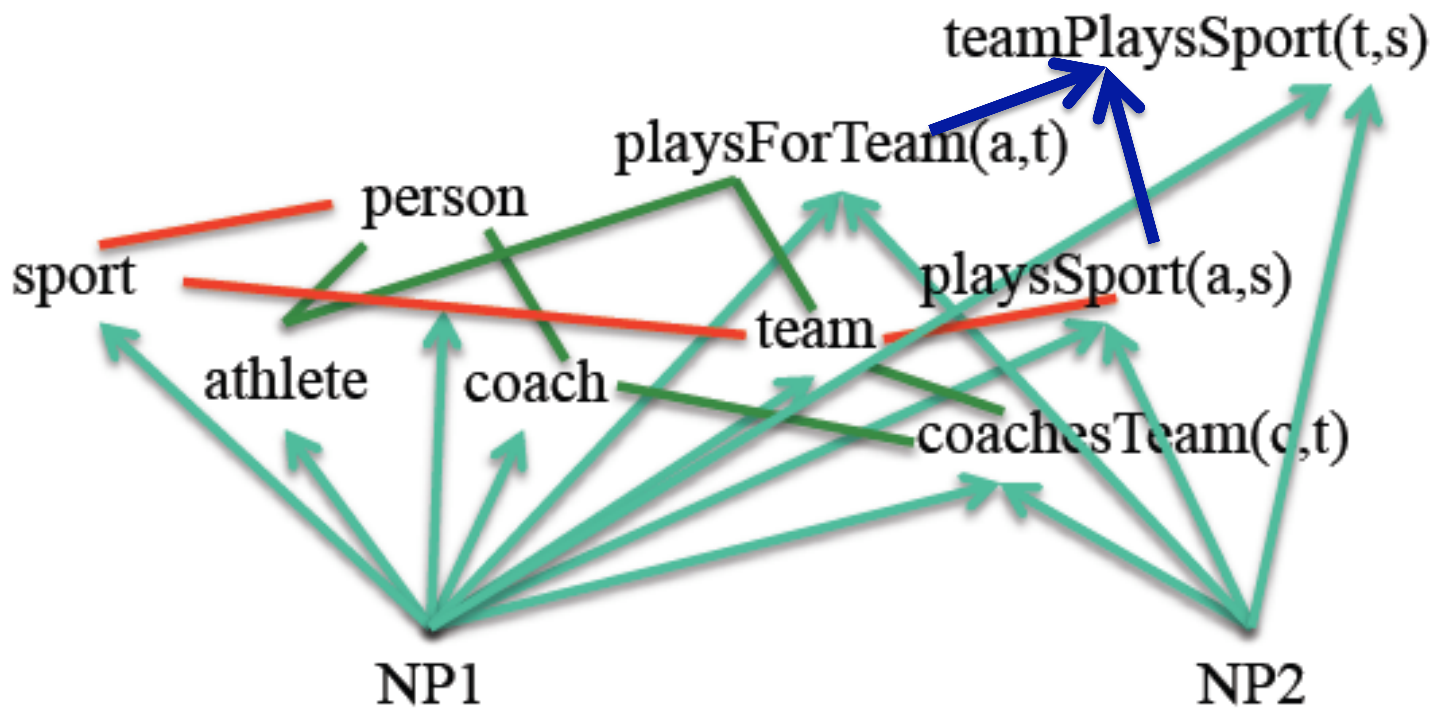
0.70 teamPlaysSport(?x,basketball) ← playsAgainst(?x,pistons)

0.64 teamPlaysSport(?x,?y) ← playsAgainst(?x ?z), teamPlaysSport(?z,?y)

…

Learned Probabilistic Horn Clause Rules

0.81 teamPlaysSport(?x,?y) ← playsForTeam(?x,?z), playSport(?z,?y)



Summary: what we have to work from

Data/Knowledge:

- KB with 10^4 extracted beliefs, .85-.90 accurate
- statistics on co-occurrence frequency of
  - 10^5 NPs x 10^5 contexts

Learning algorithms

- coupled semi-supervised learning of name entity extractors, relation extractors
- learning probabilistic first-order horn clauses

Homework 1 : due next thursday

- get co-occurrence data:10^5 NP's x 10^5 Contexts
- look at http://rtw.ml.cmu.edu/wsdm10_online/
  – labeled data: "Instances Promoted by Meta-Bootstrap Learner"
- do something interesting, prepare a 2 slide, 3 min presentation

Example:

learn to classify <NP,Context>, or just NP's as city, person, emotion, …

– supervised, semi-supervised, unsupervised, …

Projects Ideas 1

- Add a morphology-based entity extractor
  - Omalinski is probably a person
  - yet another redundant information source
- Ultra High-dimensional training
  - each noun phrase as bag of 10^6 contexts
  - each context as bag of 10^5 noun phrases
- Add first self-reflection capability
  - where is my performance weakest?
  - what should I do next?

Projects Ideas 2

- Better rule learning algorithm
  - learning from positive data only?
  - accuracy estimates based on resampling?

- Prep phrase attachment
  - How can KB and background statistics be used?

- Co-reference resolution
  - IBM versus Int.Bus.Mach versus it

Projects Ideas 3

- Active learning
  - what questions should I ask in today's 5 min session with a human?
  - what new data should I download?

- Temporal scoping
  - How can we determine *when in time* a fact holds?
  - Consider earliest web page containing the fact?
  - Date web pages?
  - Read the temporal scope?