NYU at Cold Start 2015: Experiments on KBC with NLP Novices

Yifan He Ralph Grishman
Computer Science Department
New York University

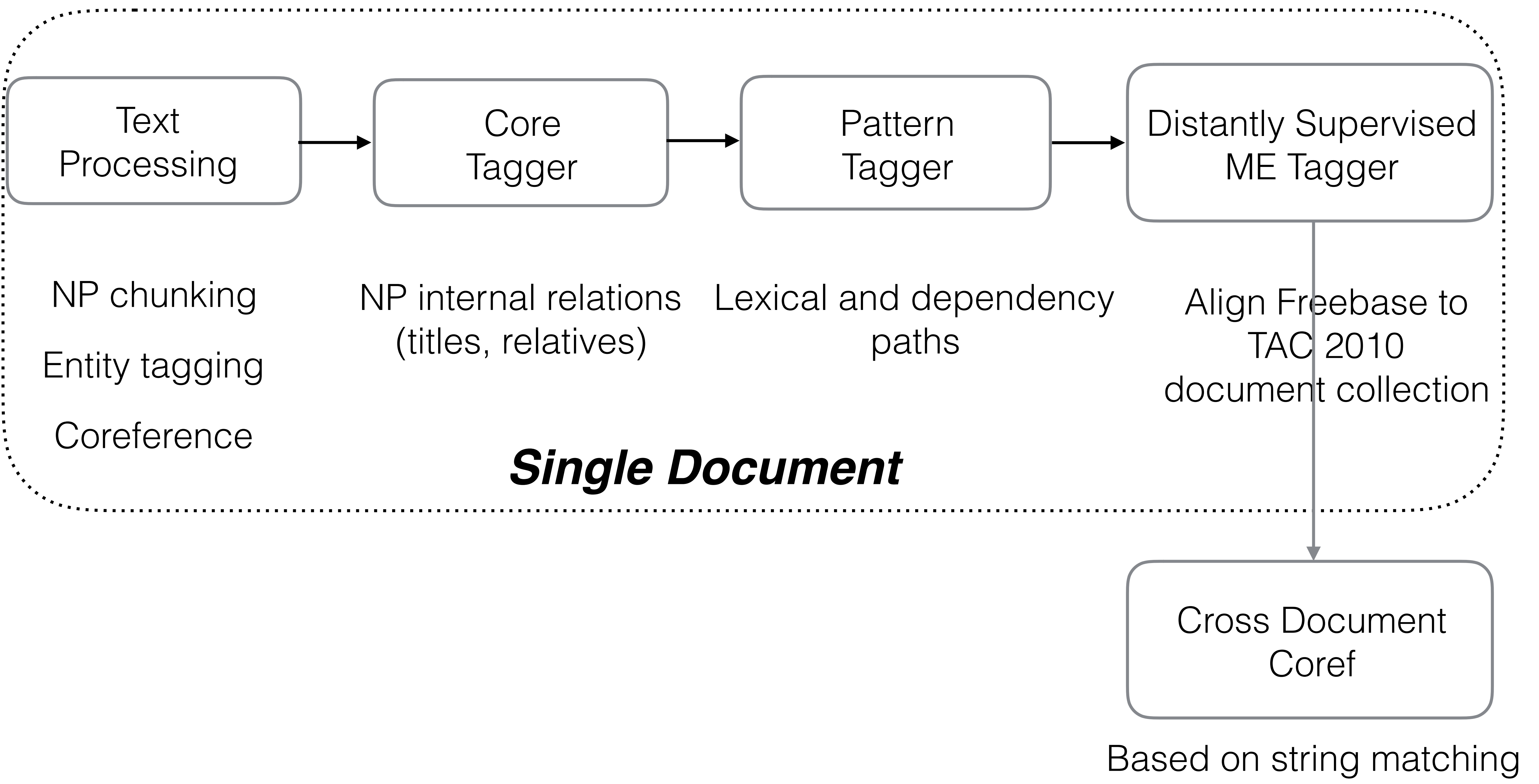
The KBP Cold Start Task and Common Approaches

- The KBP Cold Start task builds a knowledge base from scratch using a given document collection and a predefined schema for the entities and relations
- Common approaches
  - Hand-written rules (Grishman and Min, 2010)
  - Supervised relation classifiers
    - Weakly supervised classifiers: distant supervision (Mintz et al., 2009; Surdeanu et al., 2012), active learning / crowd sourcing (Angeli et al., 2014)

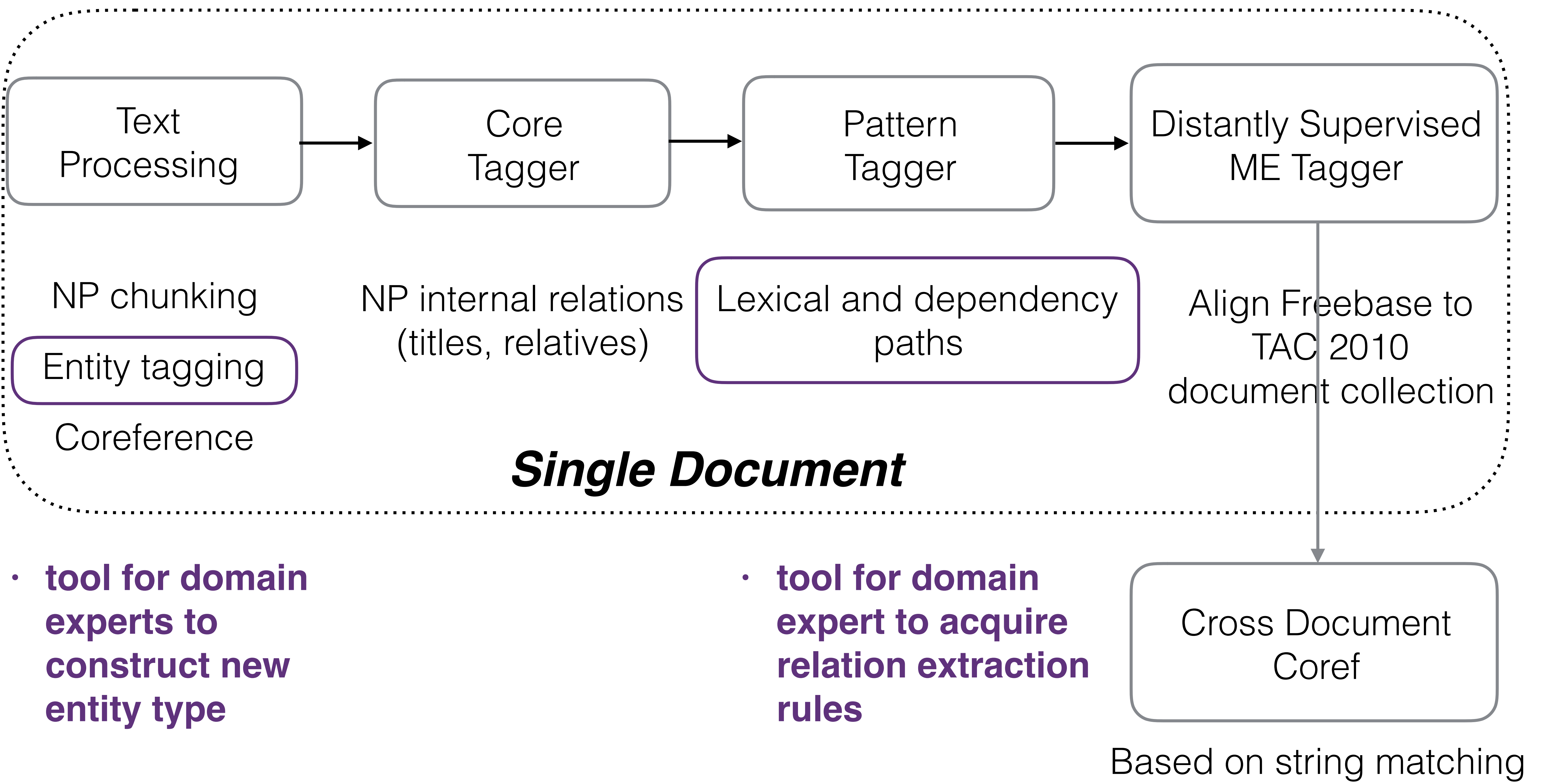
Focus this year: NLP Novices

- Current approaches often require NLP expertise
  - NYU rules are tuned every summer for 7 years
  - Supervised systems: annotation and algorithm design
  - Crowdsourcing: secret documents?
- **Can a domain expert construct an in-house knowledge base from scratch, by herself, (using tools)?**

NYU Cold Start Pipeline



NYU Cold Start Pipeline

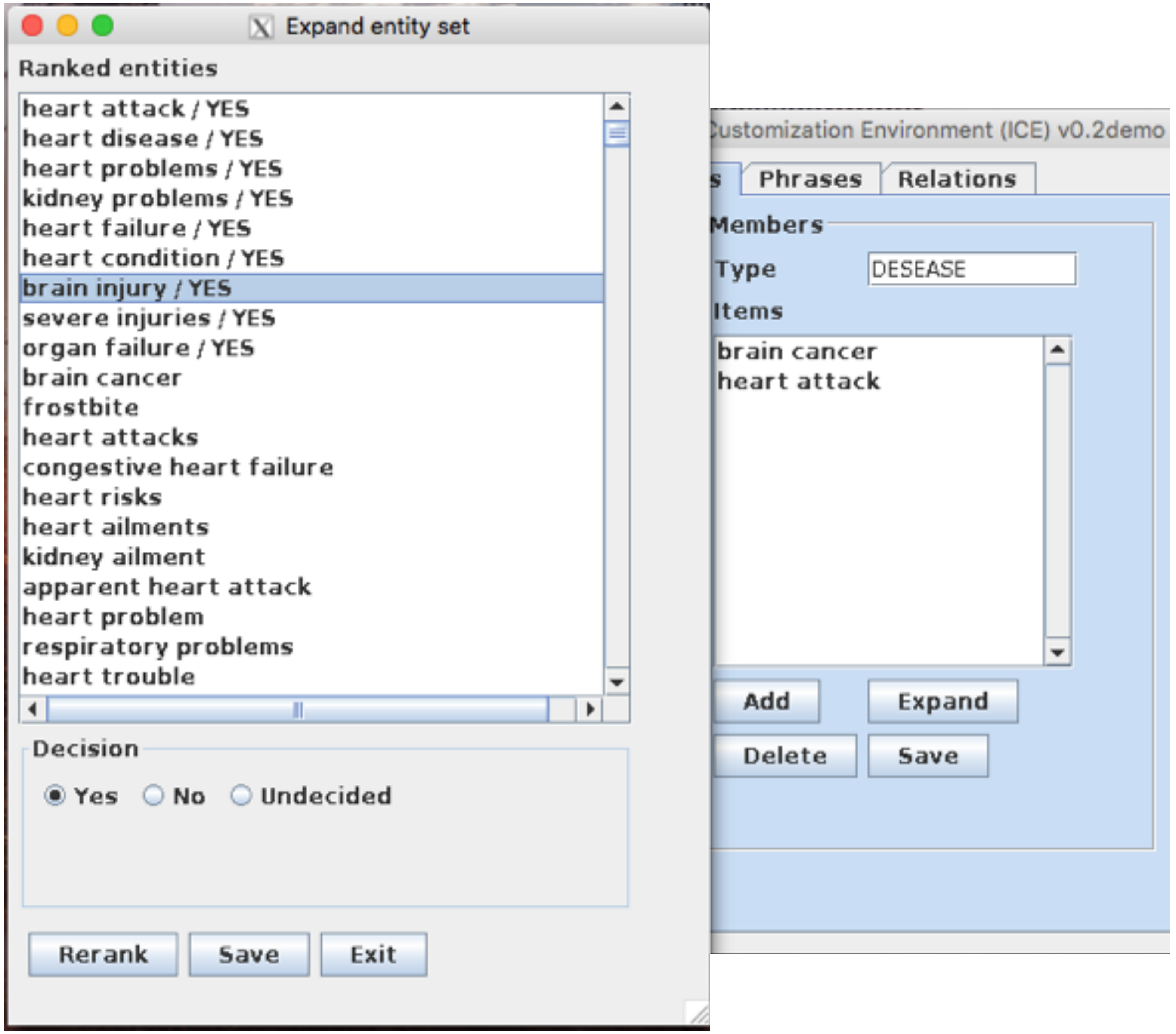


Entity Type and Relation Construction with ICE

- ICE [Integrated Customization Environment for Information Extraction]
    - easy tool for non-NLP experts to rapidly build customized IE systems for a new domain
- Entity set construction
- Relation extraction

Constructing Entity Sets

- New entity class (e.g. **DISEASE** in *per:cause_of_death*) by dictionary
  - users are not likely to do a good job assembling such a list
  - users are much better at reviewing a system-generated list
- Entity set expansion: start from 2 seeds, offer more to review



Ranking Entities

- Entities are represented with context vectors
  - Contexts are dependency paths from and to the entity
  - V_{heroin}:{dobj_sell:5, nn_plant:3, dobj_seize:4, …}
  - V_{heart_attack}:{prep_from_suffer:4, prep_of_die:3, …}
- Entities ranked by distance to the cluster centroid (Min and Grishman, 2011)

Constructing Relations: Challenges

- Handle new entity types in relation (solved by entity set expansion: ICE recognizes **DISEASE** after it is built)

- Capture variations in linguistic constructions

  - **ORGANIZATION** *leader* **PERSON** vs. **ORGANIZATION** *revived under* **PERSON** *('s leadership)*

- User comprehendible rules

Rules: Dependency Path

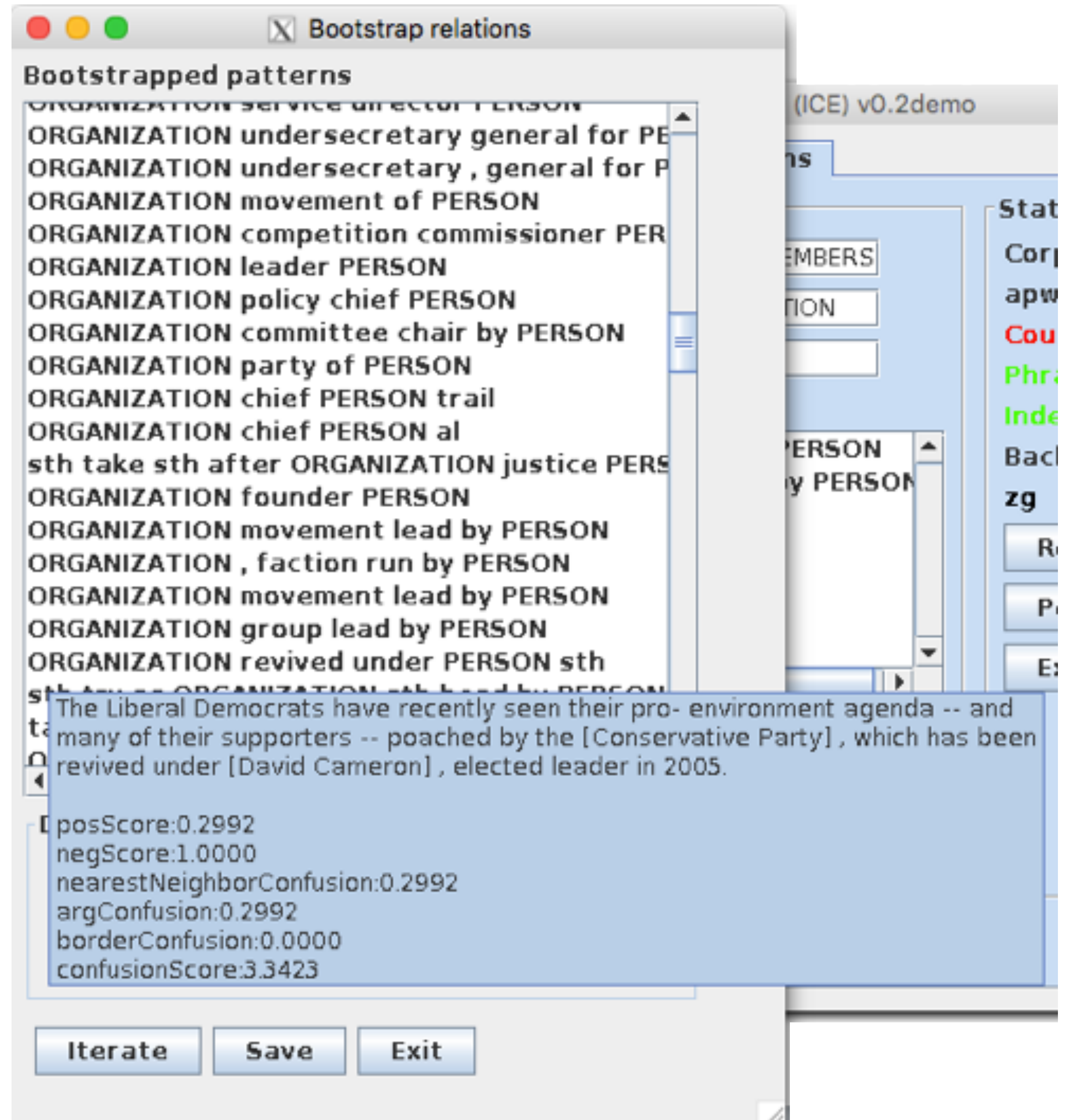
- Lexicalized dependency paths (LDPs) extractors
  - Simple, transparent approach; no feature engineering
  - Straightforward for bootstrapping
  - Most important component in NYU's slot-filling / cold start submissions (Sun et al. 2011; Min et al. 2012)

LDP
ORGANIZATION — dobj-1:revived:prep_under — PERSON

Can user understand this?

Comprehendible Rules: Linearized LDPs

- Linearize LDP into English phrases
  - User reviews linearized English phrases
  - Based on word order in original sentence
  - Insert syntactic elements for fluency: indirect objects, possessives etc.
  - Lemmatize words except passive verbs



Bootstrapping: Finding Varieties in Rules

- Dependency path acquisition with the classical (active) Snowball bootstrapping (Agichtein and Gravano, 2000)
- Algorithm skeleton

ORGANIZATION *leader* **PERSON**

Conservative_Party:Cameron

ORGANIZATION *revived under* **PERSON**

Microsoft:Nadela

ORGANIZATION *ceo* **PERSON**

1. User provide seeds
2. Collect arguments from seeds
3. New paths for review
4. Iterate

Experiments

- Entity set expansion and relation bootstrapping on Gigaword AP newswire 2008 data
  - Construct DISEASE entity type
  - Bootstrap all relations, only using seeds from slot descriptions
- **CoreTagger**: only use the core tagger which tags NP internal relations
- **Setting 1**: 5 iterations of bootstrapping, review 20 instances per iteration - 553 dependency path rules
- **Setting 2**: 5 iterations of bootstrapping, review as many phrases as possible, bootstrap with coreference (Gabbard et al., 2011) - 1,559 dependency path rules
- "**Proteus**": NYU submission that uses 1,402 dependency patterns, 2,495 lexical patterns, and an add-on distantly supervised relation classifier

Experiments

- Entity set expansion and relation bootstrapping on Gigaword AP newswire 2008 data
  - Construct DISEASE entity type
  - Bootstrap all relations, only using seeds from slot descriptions
- **CoreTagger**: only use the core tagger which tags NP internal relations
- **Setting 1**: 5 iterations of bootstrapping, review 20 instances per iteration - dependency path rules

~20 min per relation

- **Setting 2**: 5 iterations of bootstrapping, review as much as possible, bootst with coreference (Gabbard et al., 2011) - 1,559 dependency path rules

~1 hr per relation

- “**Proteus**”: NYU submission that uses 1,402 dependency patterns, 2,495 lexical patterns, and an add-on distantly supervised relation classifier

7 summers

Results: Hop0

	P	R	F
CoreTagger	0.71	0.06	0.11
CoreTagger +Setting1	0.44	0.08	0.13
CoreTagger +Setting2	0.54	0.13	0.21
CoreTagger +Proteus	0.46	0.25	0.32

TAC 2014 Evaluation Data; Proteus = Patterns + Fuzzy Match + Distant Supervision

Results: Hop0+Hop1

	P	R	F
CoreTagger	0.47	0.04	0.07
CoreTagger +Setting1	0.34	0.05	0.08
CoreTagger +Setting2	0.37	0.08	0.13
CoreTagger +Proteus	0.31	0.20	0.24

TAC 2014 Evaluation Data; Proteus = Patterns + Fuzzy Match + Distant Supervision

Summary

- Pilot experiments on bootstrapping a KB constructor from scratch using an open-source tool
  - Builds high-precision/modest recall KBs
  - Friendly to domain experts who are not familiar with NLP: user only reviews plain English examples
  - Builds rule-based interpretable models for both entity and relation recognition

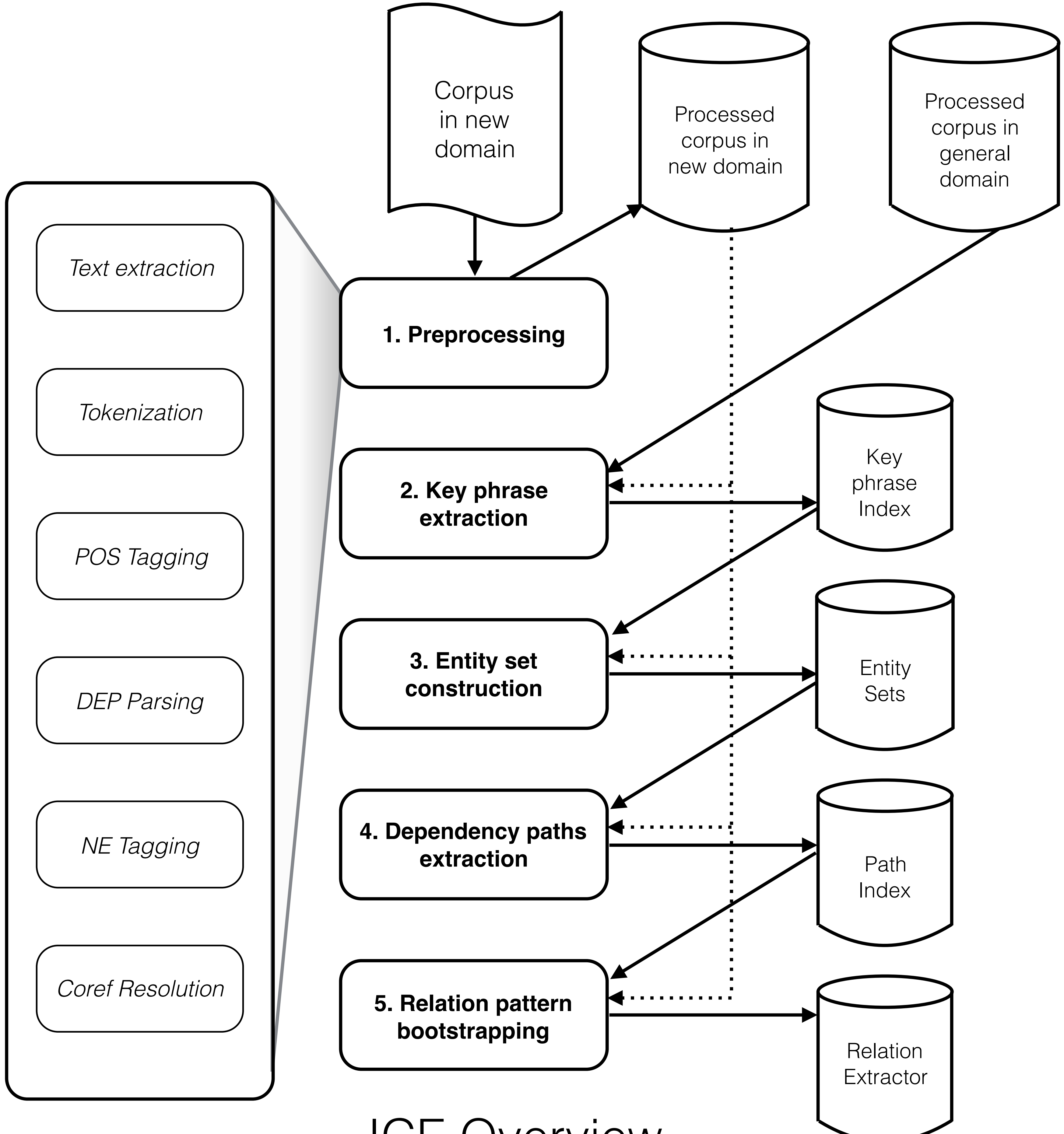
More To Be Done

- Better annotation instance selection
    - So that the casual user can perform similarly to a serious user
- More expressive rules beyond dependency paths
    - Event extraction
- Leverage existing KB

Thank you

http://nlp.cs.nyu.edu/ice

http://github.com/rgrishman/ice



ICE Overview

/event/disaster/structures_damaged

Filter options: ☐ Include deleted links | Timestamp YYYY-MM-DD... to YYYY-MM-DD...
View options: ☐ Sort oldest to newest | ☐ Show full timestamp | ☐ Show full attribution

Links

	Subject	Predicate	Object/Value
1	/m/0gg9kfr 2011 Christchurch earthquake	/event/disaster/structures_damaged	/m/0j_2yw_ St Luke's Church, Christchurch
2	/m/0gg9kfr 2011 Christchurch earthquake	/event/disaster/structures_damaged	/m/0gg7hn1 Hotel Grand Chancellor, Christchurch
3	/m/0gg9kfr 2011 Christchurch earthquake	/event/disaster/structures_damaged	/m/0by116z Christchurch Hospital
4	/m/0qtwtw9 Chelyabinsk Event	/event/disaster/structures_damaged	/m/0r944hl Ice Palace "Ural Lightning"
5	/m/0qtwtw9 Chelyabinsk Event	/event/disaster/structures_damaged	/m/0qzqcvy Chelyabinsk Zinc Factory
6	/m/0qtwtw9 Chelyabinsk Event	/event/disaster/structures_damaged	/m/0qtx4gt Chelyabinsk Drama Theatre
7	/m/0qtwtw9 Chelyabinsk Event	/event/disaster/structures_damaged	/m/064pnfg Traktor Ice Arena
8	/m/0j0z2w4 Port Said Stadium disaster	/event/disaster/structures_damaged	/m/0b72l9 Port Said Stadium
9	/m/0gh6mkc 2011 Tōhoku earthquake and tsunami	/event/disaster/structures_damaged	/m/02vk_7d Fukushima Daini Nuclear Power Plant
10	/m/0gh6mkc 2011 Tōhoku earthquake and tsunami	/event/disaster/structures_damaged	/m/02vkzy2 Fukushima Daiichi Nuclear Power Plant
11	/m/0b4mlj Katowice Trade Hall roof collapse	/event/disaster/structures_damaged	/m/02r05rb Katowice International Fair
12	/m/01v8cd Summerland disaster	/event/disaster/structures_damaged	/m/05bgrl4 Summerland Leisure Centre
13	/m/0dc3pc Royal Suspension Chain Pier	/event/disaster/structures_damaged	/m/0dc3pc Royal Suspension Chain Pier
14	/m/05252dm Tay Bridge disaster	/event/disaster/structures_damaged	/m/04zjqhp The Tay Bridge
15	/m/098sht Buncefield fire	/event/disaster/structures_damaged	/m/098sp5 Buncefield oil depot
16	/m/0d0vp3 September 11 attacks	/event/disaster/structures_damaged	/m/09w3b The Pentagon
17	/m/0807k3 1983 United States Senate bombing	/event/disaster/structures_damaged	/m/07vth United States Capitol
18	/m/01y23_ 16th Street Baptist Church bombing	/event/disaster/structures_damaged	/m/0bf9_v 16th Street Baptist Church
19	/m/0244k9 MGM Grand fire	/event/disaster/structures_damaged	/m/033vpy MGM Grand Las Vegas
20	/m/053zwd 1996 Garley Building fire	/event/disaster/structures_damaged	/m/05bgrkg Garley Building
21	/m/07hxss 1992 Windsor Castle fire	/event/disaster/structures_damaged	/m/0chgsm Windsor Castle
22	/m/0b_94y Whiskey Au Go Go fire	/event/disaster/structures_damaged	/m/05bgrnw Whiskey Au Go Go
23	/m/02vnpxc Uphaar Cinema fire	/event/disaster/structures_damaged	/m/05bgrjk Uphaar Cinema
24	/m/0b27k1 Dee Bridge disaster	/event/disaster/structures_damaged	/m/0cfgmk Old Dee Bridge

Entity Set Expansion/ Ranking

- In each iteration, present the user with ranked entity list, ordered by the distance to the "positive centroid" (Min and Grishman, 2011):

$$c = \frac{\sum_{p \in P} p}{|p|} - \frac{\sum_{n \in N} n}{|n|}$$

- where c is the positive centroid, P is the set of positive seeds (initial seeds and entities accepted by user), and N is the set of negative seeds (entities rejected by user)

- Update centroid for k iterations

Entity Representation

- Represent each phrase with a context vector, where contexts are dependency paths from and to the phrase
  - DRUGS share *dobj*(sell, X) and *dobj*(seize, X) contexts
  - DISEASE share prep_of(die, X) and prep_from(suffer) contexts
- Examples: count vectors of dependency contexts
  - V_{heroin}:{dobj_sell:5, nn_plant:3, dobj_seize:4, …}
  - V_{heart_attack}:{prep_from_suffer:4, prep_of_die:3, …}
- Features weighted by PMI; word embedding on large data sets for dimension reduction

Entity Representation II

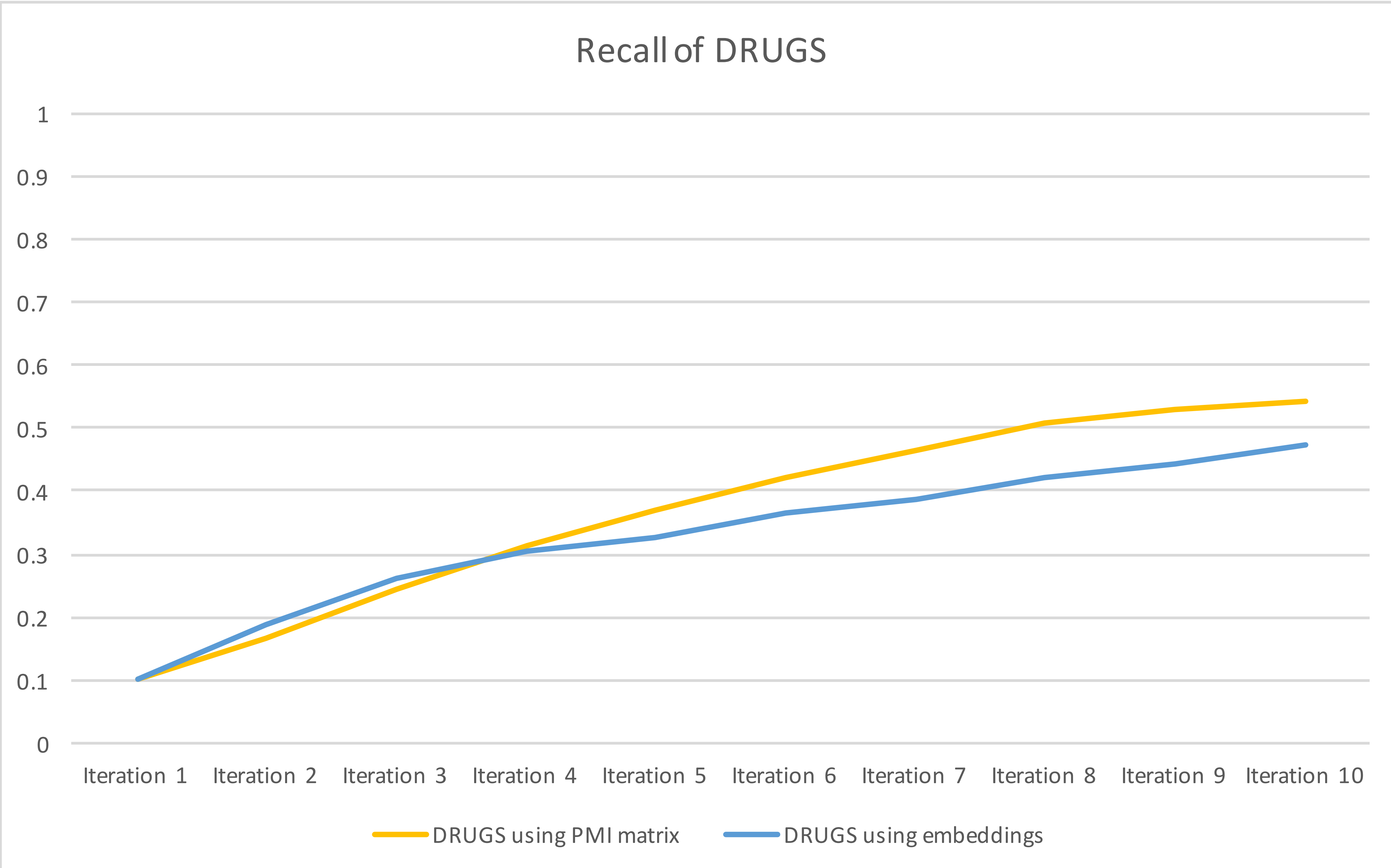
- Using raw vectors cannot provide live response
- Dimension reduction via word embeddings
- Skip-gram model with negative sampling, using dependency context (Levy and Goldberg, 2014a)
- Equivalent of factorization of the original* feature matrix (Levy and Goldberg, 2014b)

* shifted; PPMI instead of PMI0

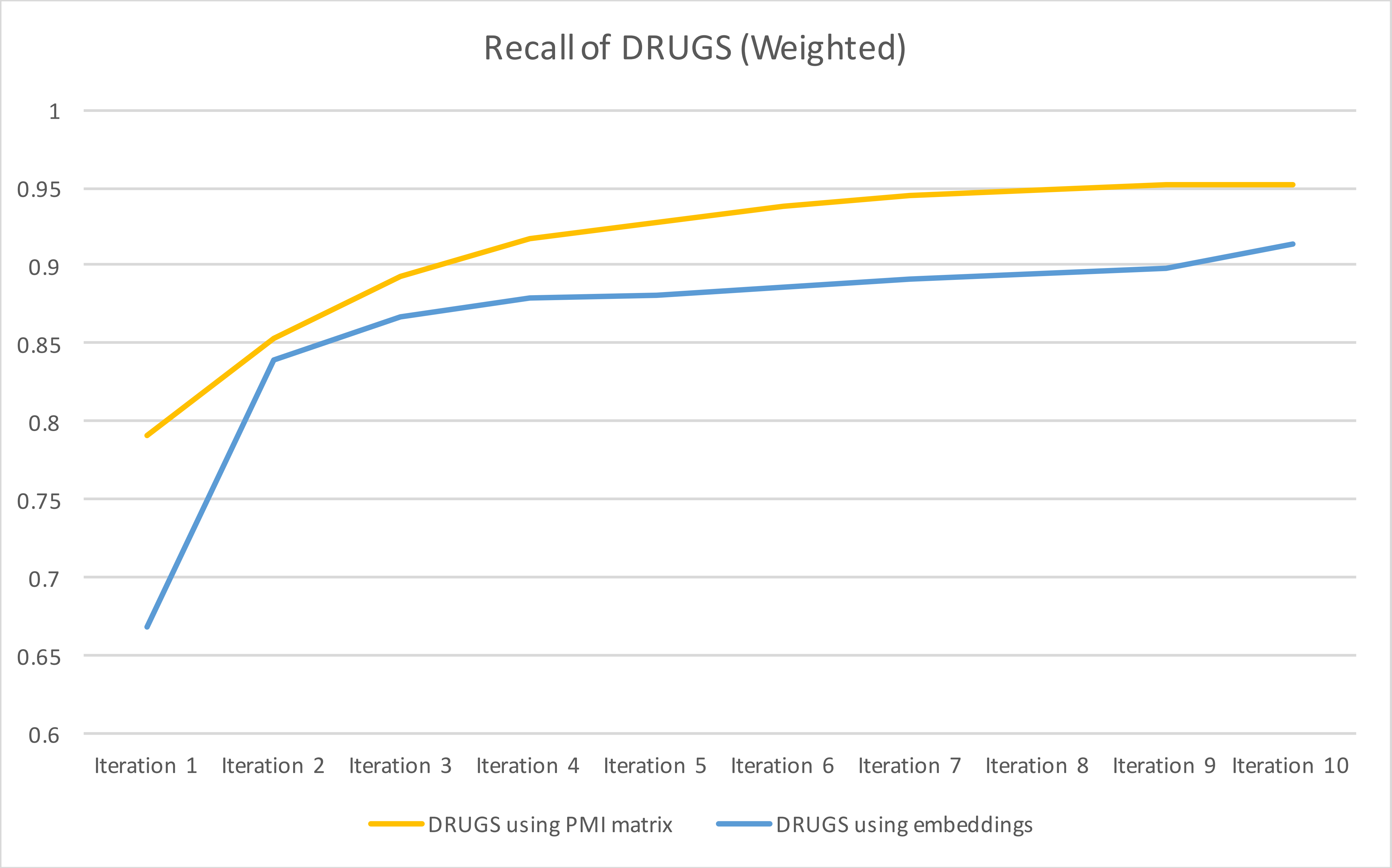
Experiment of Entity Set Expansion

- Finding Drugs in Drug Enforcement Agency news releases
- 10 iterations, review 20 entity candidates per iteration
- Measure recall on a pre-compiled list of 181 drug names from 2,132 key phrases
- DISEASES: ICE 129 diseases; Manual 19 diseases

Constructing **Drugs** Type

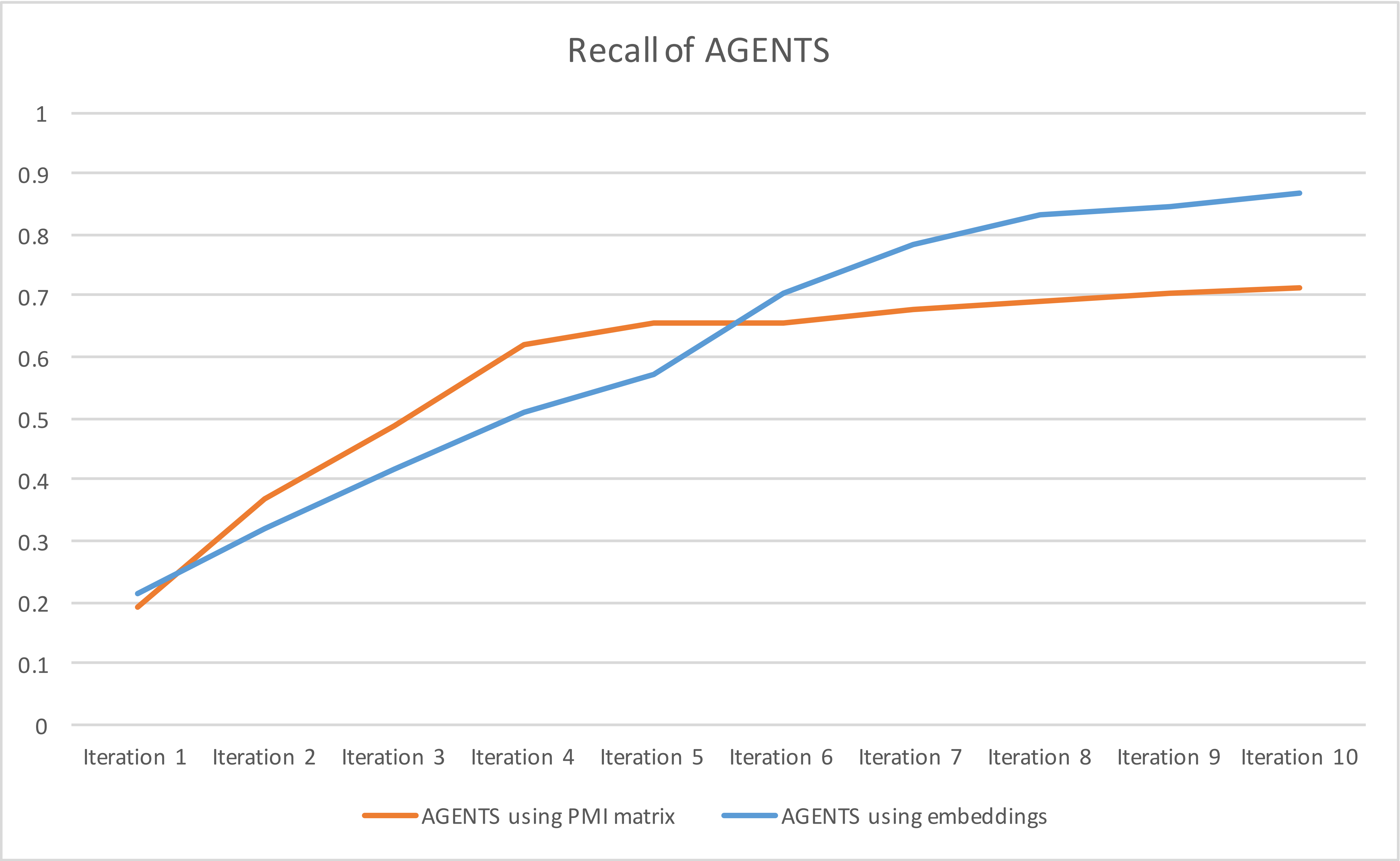


Constructing **Drugs** Type (Weighted Result)



- Recall score weighted by frequency of entities

Results - Agents



- 84 positive examples from 2,132 candidates

Results: Hop0 - w/ FM

	P	R	F
CoreTagger	0.71	0.06	0.11
CoreTagger +Setting1	0.44	0.08	0.13
CoreTagger +Setting2	0.41	0.11	0.18
CoreTagger +Proteus	0.46	0.25	0.32

TAC 2014 Evaluation Data; Proteus = Patterns + Fuzzy Match + Distant Supervision

Results: Overall - w/ FM

	P	R	F
CoreTagger	0.47	0.04	0.07
CoreTagger +Setting1	0.34	0.05	0.08
CoreTagger +Setting2	0.31	0.10	0.15
CoreTagger +Proteus	0.31	0.20	0.24

TAC 2014 Evaluation Data; Proteus = Patterns + Fuzzy Match + Distant Supervision

Fuzzy dependency path match for small rule set

- Improve recall for small rule sets
  - Also tested in our 2015 KBP Cold Start submission
- Match two LDPs with edit distance on dependency chains
  - Weight of edit operations set by grid search on dev set (substitution: 0.8, insertion: 1.2, deletion: 0.3; feature-based see paper)
  - Substitution cost determined by word similarity based on word embeddings

Fuzzy dependency path match-based extraction: example

	nsubj-1:sell	dobj:prescription	nn-1:END$
dsubj:END$		0.3	0
nsubj-1:ditribute	0.28*0.8		

Edit costs
substitution: 0.8
insert: 1.2
delete: 0.3

$$cost = \frac{weightedDistance}{|rule|}$$

$$= \frac{0.28 * 0.8 + 0.3}{3}$$

$$= 0.17$$

Official Run Results

	NestedNames+Pattern+DS+FM			Pattern+DS		
	P	R	F	P	R	F
Hop0	0.44	0.20	0.27	0.51	0.18	0.27
Hop1	0.06	0.09	0.07	0.15	0.09	0.11
MicroAvg	0.17	0.15	0.16	0.30	0.14	0.20
MacroAvg			0.18			0.17

Main goal: testing the fuzzy match paradigm
False positives on NIL slots from Fuzzy Match in Hop 0 was penalized heavily in Hop1