Weakly Supervised Named Entity Recognition

Joseph Birkner
Seminar for Information Extraction
CIS – LMU
WS 16/17

Table of Contents

1.1 Premise

- Human annotation (Supervision) is expensive.

1.1 Premise

- Human annotation (Supervision) is expensive.
- Un-annotated data is ~~cheap~~ free

1.1 Premise

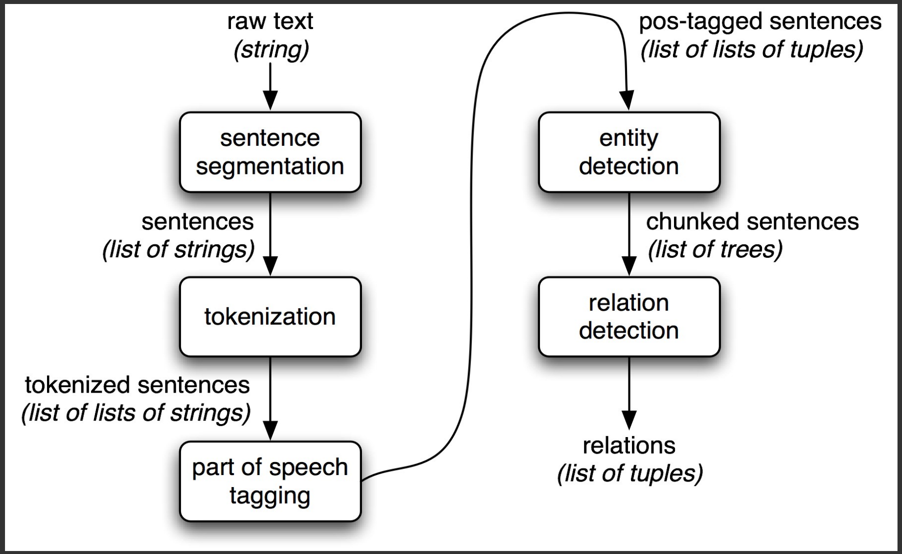
- Human annotation (Supervision) is expensive.
- Un-annotated data is ~~cheap~~ free

$\longrightarrow$ We want systems that

  - Maximize use of unannotated data
  - Minimize need for human input (Supervision)

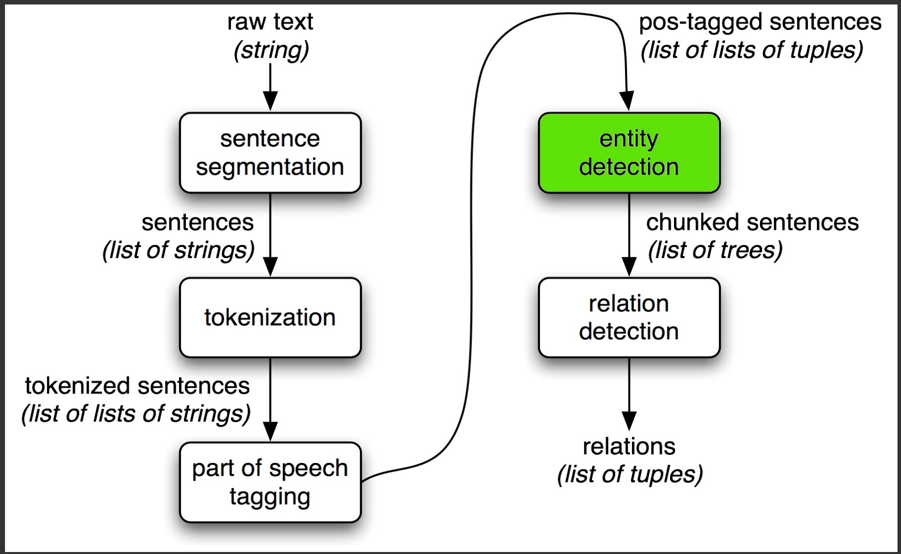
- Learn with a minimum of human supervision!
  - Semi-Supervised Learning
  - Like Skynet!

1.2 What is this talk about?



(nltk.org)

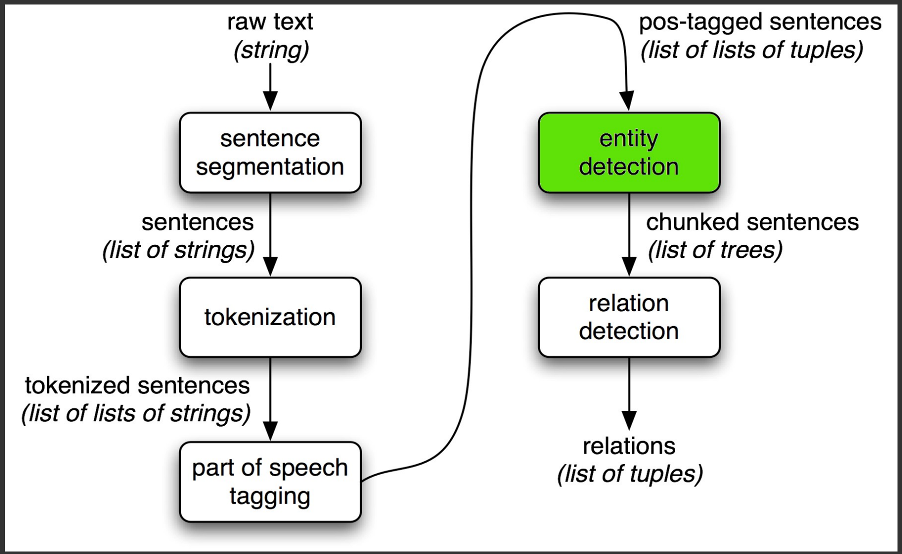
1.2 What is this talk about?



(nltk.org)

- Will not talk about segmentation/tagging.

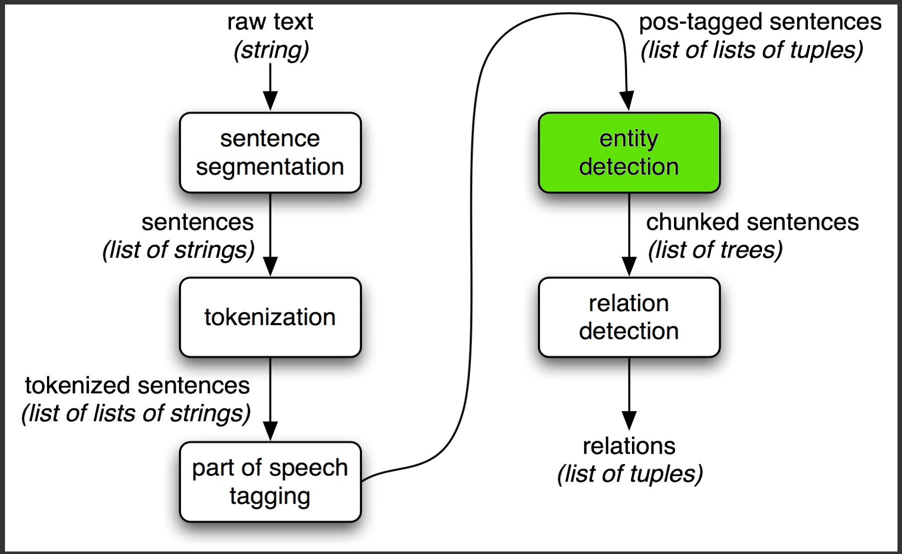
1.2 What is this talk about?



(nltk.org)

- Will not talk about segmentation/tagging.
- Will only talk about models that work for both recognition (detection) and classification!

1.2 What is this talk about?



(nltk.org)

- Will not talk about segmentation/tagging.
- Will only talk about models that work for both recognition (detection) and classification!
- Will not talk about active (reinforcement) learning.
  - So no Skynet

1.2 What is this talk about?

2013
Lightly Supervised NER
Distantly Supervised NER

2014
Distantly Supervised NER
Distantly Supervised NER

2015
Lightly Supervised NER
Lightly Supervised NER

Now
Weakly Supervised NER

1.2 What is this talk about?

- This raises some questions:

Weak Supervision

Semi-Supervision

Light Supervision

Distant Supervision

1.2 What is this talk about?

- This raises some questions:

(Nadeau, 2007)

Weak Supervision = Semi-Supervision

Light Supervision

Distant Supervision

1.2 What is this talk about?

- This raises some questions:

(Nadeau, 2007)

Weak Supervision = Semi-Supervision

Light Supervision

(Sanchez et al., 2011)

Distant Supervision

(Mintz et al., 2009)

1.2 What is this talk about?

- Semi-Supervised Learnig + NER?

1.2 What is this talk about?

- Semi-Supervised Learnig + NER?
- PU Learning
  - Set of labeled (positive) examples P
  - Set of unlabeled examples U
  - Optional: Negative Examples

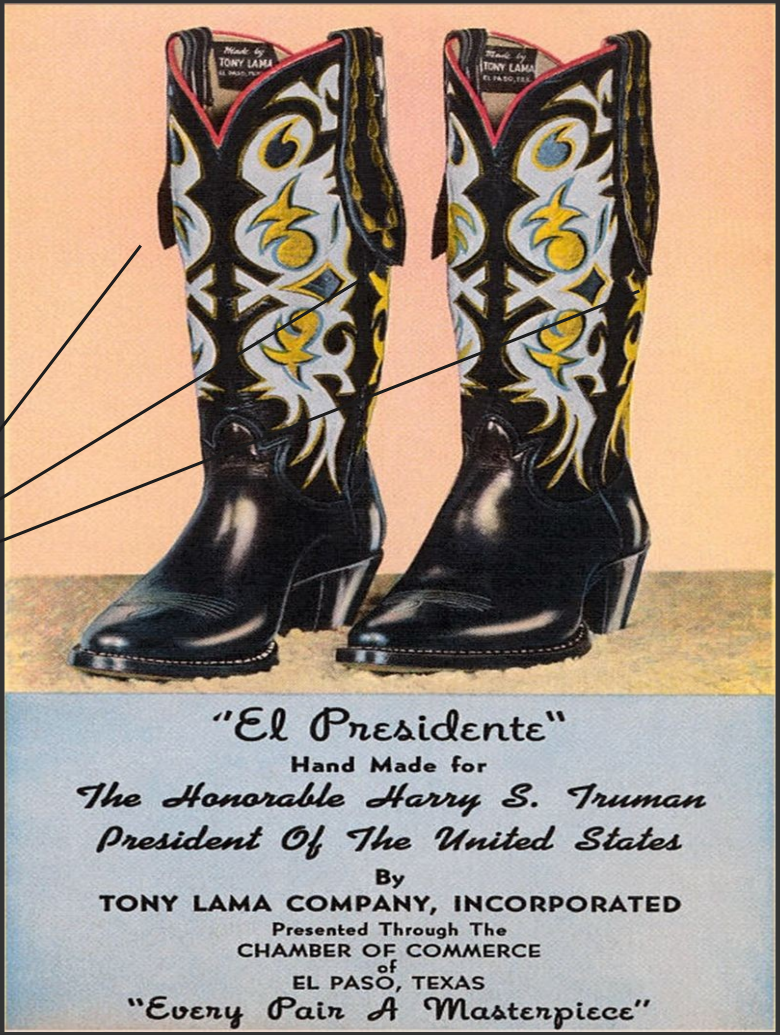
1.2 What is this talk about?

- Semi-Supervised Learnig + NER?
- PU Learning
  - Set of labeled (positive) examples P
  - Set of unlabeled examples U
  - Optional: Negative Examples
- Bootstrapping

1.3 Bootstrapping

(Wikimedia Commons)

Bootstraps



1.3 Bootstrapping

- Fairy Tale
- Unscientific
- Defies Newton

- Does this apply to Semi-Supervised Learning?
- What can we possibly learn from unannotated data?

(Theodor Hosemann / Wikimedia Commons)

1.3 Bootstrapping

- Unlabeled text provides joint probability distributions

```
Labeled:  Grace Hopper and her husband divorced in 1945.
Model:        and, husband: very likely in close context
Unlabeled:    Neil Patrick Harris and husband David Burtka
              often share adorable snapshots
New context: actor, played: very likely in close context
```

1.3 Bootstrapping

- Unlabeled text provides joint probability distributions

```
Labeled:  Grace Hopper and her husband divorced in 1945.
Model:        and, husband: very likely in close context
Unlabeled:    Neil Patrick Harris and husband David Burtka
              often share adorable snapshots
New context: actor, played: very likely in close context
```

- Unlabeled text provides context tokens

```
Labeled:  Grace Hopper and her husband divorced in 1945.
Model:        “X and __ husband”
Unlabeled:    Neil Patrick Harris and husband David Burtka
              often share adorable snapshots
New context: actor X played
```

1.3 Bootstrapping

- Unlabeled text provides joint probability distributions

```
Labeled:  Grace Hopper and her husband divorced in 1945.
Model:       and, husband: very likely in close context
Unlabeled:   Neil Patrick Harris and husband David Burtka
             often share adorable snapshots
New context: actor, played: very likely in close context
```

- Unlabeled text provides context tokens

```
Labeled:  Grace Hopper and her husband divorced in 1945.
Model:       “X and __ husband”
Unlabeled:   Neil Patrick Harris and husband David Burtka
             often share adorable snapshots
New context: “actor X played”
```

- We can extract new context from unannotated data!

1.3 Bootstrapping

- Unlabeled text provides joint probability distributions

```
Labeled:  Grace Hopper and her husband divorced in 1945.
Model:        and, husband: very likely in close context
Unlabeled:    Neil Patrick Harris and husband David Burtka
              often share adorable snapshots
New context: actor, played: very likely in close context
```

- Unlabeled text provides context tokens

while {...

```
Labeled:  Grace Hopper and her husband divorced in 1945.
Model:        “X and __ husband”
Unlabeled:    Neil Patrick Harris and husband David Burtka
              often share adorable snapshots
New context: “actor X played”
```

- We can extract new context from unannotated data!

2.1 Bootstrapping from Examples

- (Riloff & Jones, 1999, page 2)

Generate all candidate extraction patterns from the training corpus using AutoSlog.

Apply the candidate extraction patterns to the training corpus and save the patterns with their extractions to *EPdata*

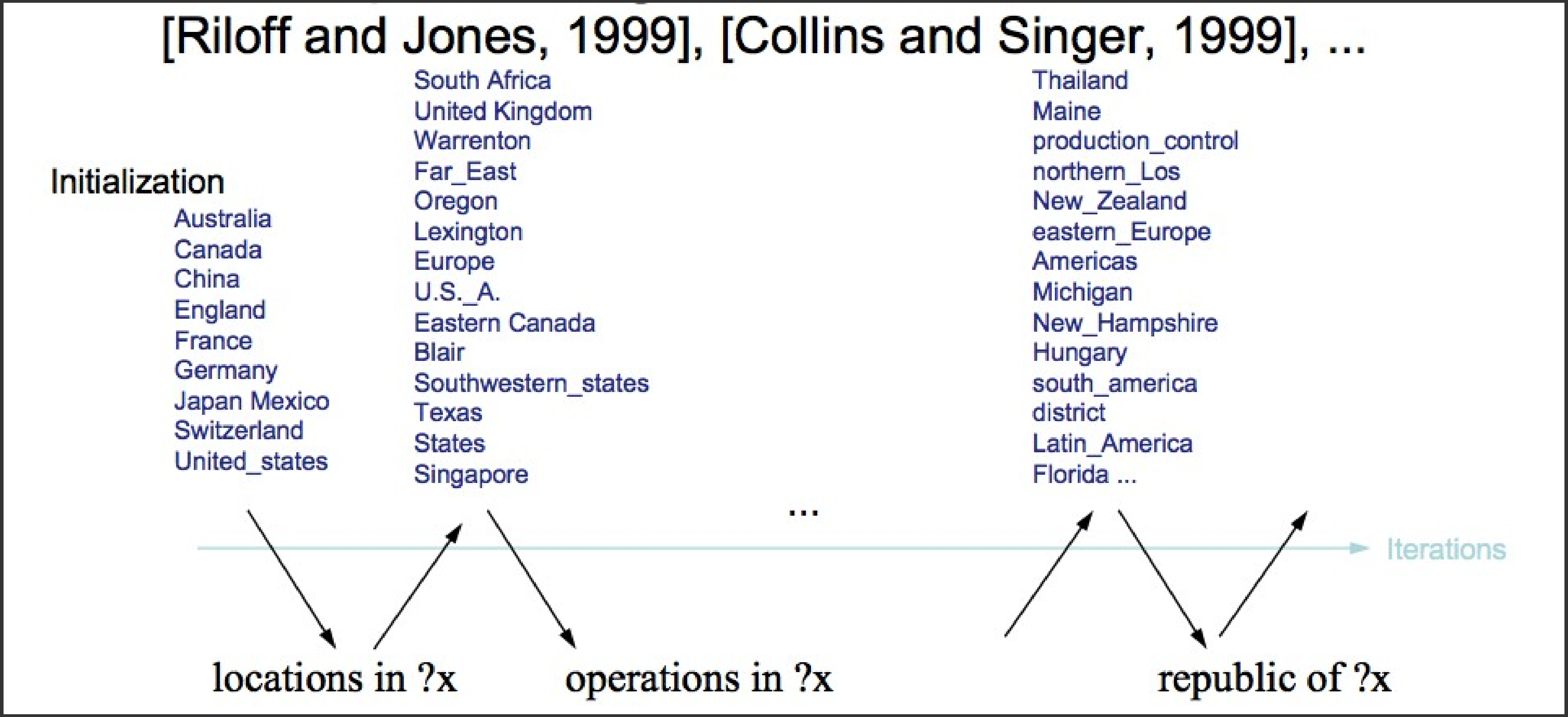
$SemLex$ = {seed_words}

Cat_EPlist = {}

MUTUAL BOOTSTRAPPING LOOP

1. Score all extraction patterns in *EPdata*.
2. *best_EP* = the highest scoring extraction pattern not already in *Cat_EPlist*
3. Add *best_EP* to *Cat_EPlist*
4. Add *best_EP*'s extractions to *SemLex*.
5. Go to step 1

2.1 Bootstrapping from Examples



(Slide from Mitchell, 2006)

2.1 Bootstrapping from Examples

- Problem: A chain is only as strong as its weakest link…

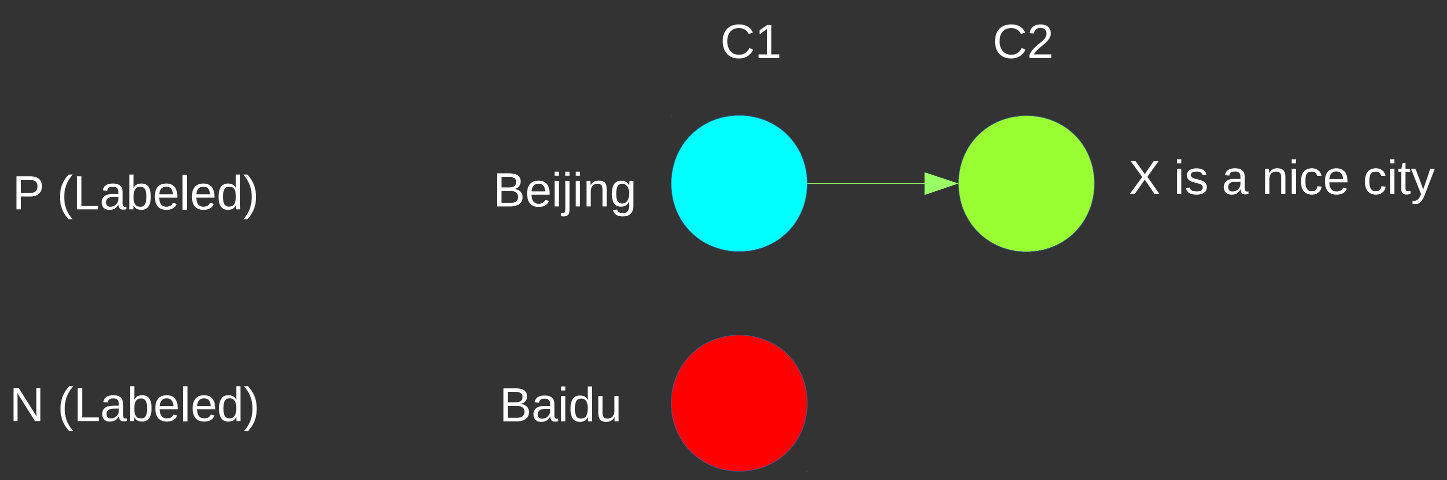
	Iter 1	*Iter 10*	*Iter 20*	*Iter 30*	*Iter 40*	*Iter 50*
Web Company	5/5 (1)	25/32 (.78)	52/65 (.80)	72/113 (.64)	86/163 (.53)	95/206 (.46)
Web Location	5/5 (1)	46/50 (.92)	88/100 (.88)	129/150 (.86)	163/200 (.82)	191/250 (.76)
Web Title	0/1 (0)	22/31 (.71)	63/81 (.78)	86/131 (.66)	101/181 (.56)	107/231 (.46)
Terr. Location	5/5 (1)	32/50 (.64)	66/100 (.66)	100/150 (.67)	127/200 (.64)	158/250 (.63)
Terr. Weapon	4/4 (1)	31/44 (.70)	68/94 (.72)	85/144 (.59)	101/194 (.52)	124/244 (.51)

Table 1: Accuracy of the Semantic Lexicons

- (Riloff & Jones, 1999, page 5)

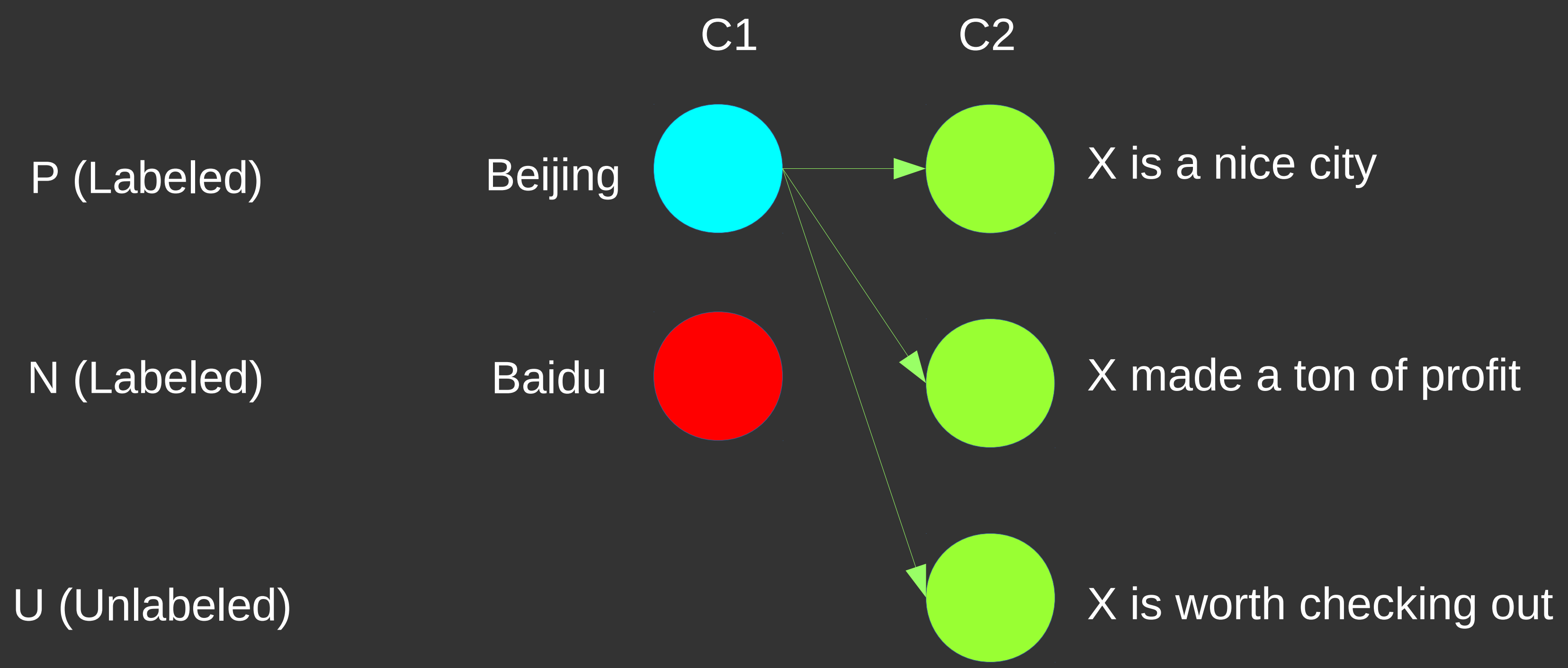
2.2 CoTraining

- (Nigam & Gani, 2000)
- Use two classifiers which should train each other!



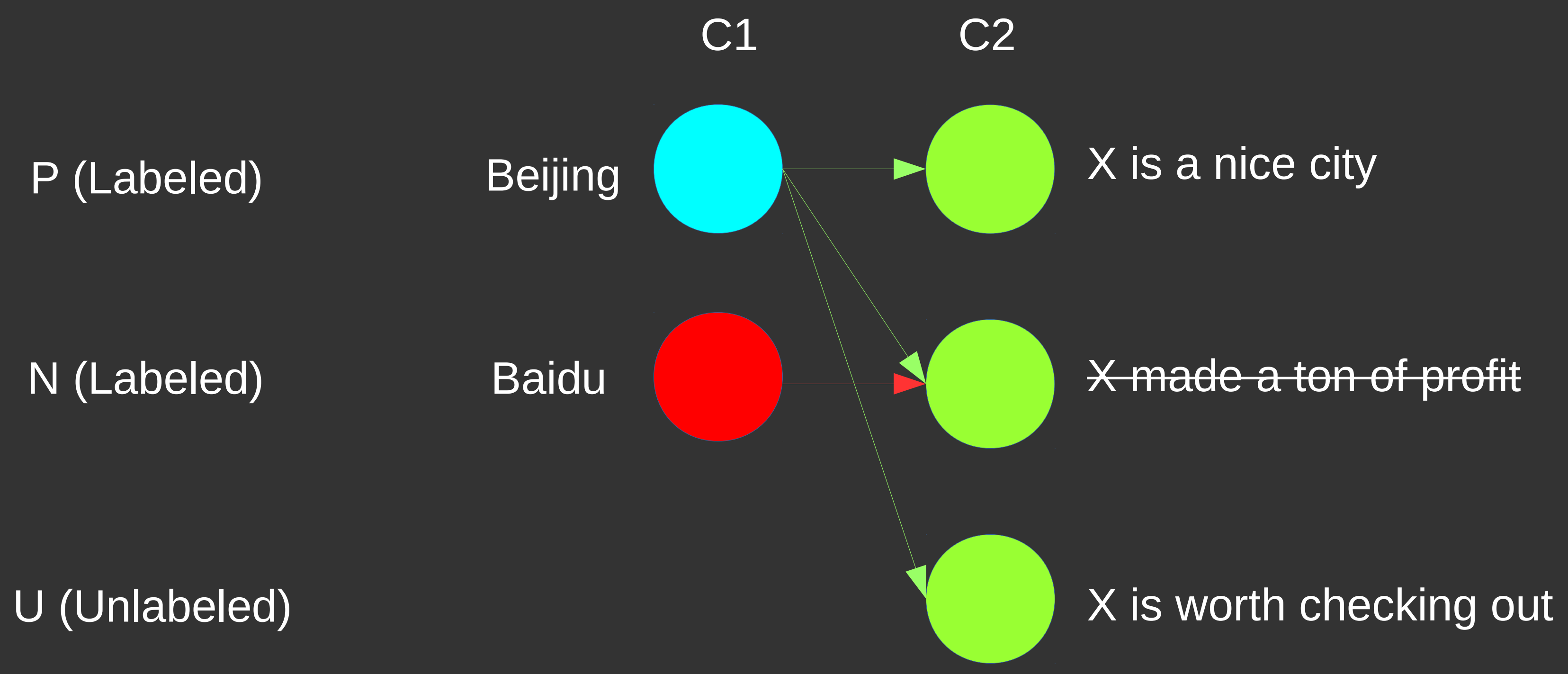
2.2 CoTraining

- (Nigam & Gani, 2000)
- Use two classifiers which should train each other!



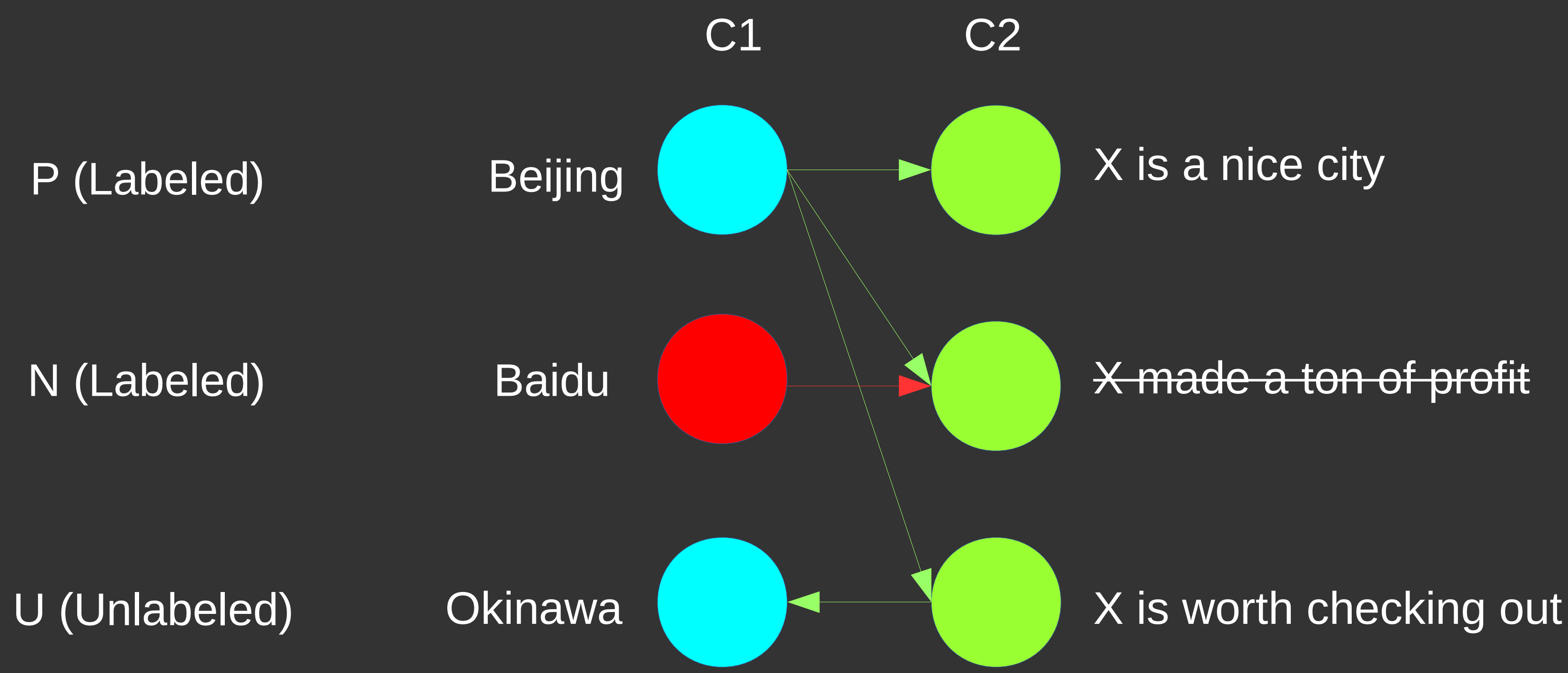
2.2 CoTraining

- (Nigam & Gani, 2000)
- Use two classifiers which should train each other!



2.2 CoTraining

- (Nigam & Gani, 2000)
- Use two classifiers which should train each other!



2.2 CoTraining

- Problems:
  - Rules are always absolutely discarded or retained
  - Discrete Classification does not allow for overlap between semantic categories

2.3 Expectation Maximization

- (Dempster, Laird, Rubin 1977)

We now present a simple characterization of the EM algorithm which can usually be applied when (2.1) holds. Suppose that $\boldsymbol{\phi}^{(p)}$ denotes the current value of $\boldsymbol{\phi}$ after p cycles of the algorithm. The next cycle can be described in two steps, as follows:

E-*step*: Estimate the complete-data sufficient statistics $\mathbf{t}(\mathbf{x})$ by finding

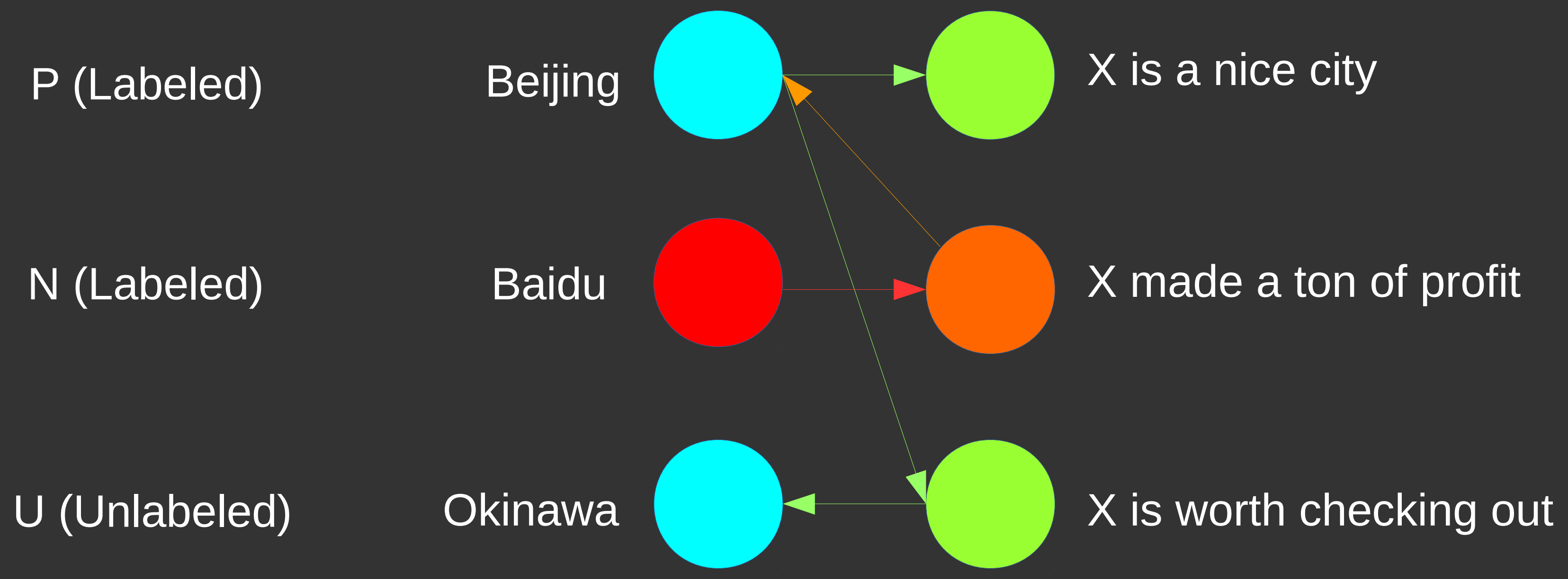
$$\mathbf{t}^{(p)} = E(\mathbf{t}(\mathbf{x}) \mid \mathbf{y}, \boldsymbol{\phi}^{(p)}). \tag{2.2}$$

M-*step*: Determine $\boldsymbol{\phi}^{(p+1)}$ as the solution of the equations

$$E(\mathbf{t}(\mathbf{x}) \mid \boldsymbol{\phi}) = \mathbf{t}^{(p)}. \tag{2.3}$$

2.4 CoEM

- (Rosie Jones, 2005): Instead of 2 classifiers, use EM
- Weak Labeling:
  - Hold a class vector over all Named Entities



2.5 Bootstrapping with Dist. Sim.

- (Pasca et al., 2005)
- Generate Patterns that consider semantic Similarity
  - Grace Hopper was born on December 9th, 1906.
  - <X> was born on {Jan, Feb, …}

2.6 Distant Supervision

- (Mintz et al. 2009)
- Originally deviced for Relation Extraction:

  “If two entities participate in a relation, any sentence that contain those two entities might express that relation”

  Knowledge Base: president(Obama, USA)

  Text: Mr. Obama is the president of the USA.

  Pattern: <X> is the president of the <Y>

- Term now used for unsupervised NER that uses a Knowledge Base for Bootstrapping
  - (Ritter 2011, Grave 2014)

2.7 Light Supervision

- (Sanchez, Bedmar, Martinez, Maqueda 2012)
- Evolved from (Riloff & Jones, 1999)
- Graph approach from CoTraining
- Weak Labeling from CoEM
- Much more fine-grained evaluation of Rules, Instances and mentions:
  - (Riloff & Jones, 1999): $score(pattern_i) = R_i * log_2(F_i)$
  - (Sanchez et al., 2012):

$$Unk(p, R_k) = \left\| \left\{ (p, e) : e \notin E_j^{t-1} \right\} \right\|$$

$$Conf(p, R_k) = \frac{Pos(p, R_k) - Neg(p, R_k)}{Pos(p, R_k) + Neg(p, R_k) + Unk(p, R_k)}$$

2.7 Light Supervision

- (Sanchez et al. 2012, page 11)

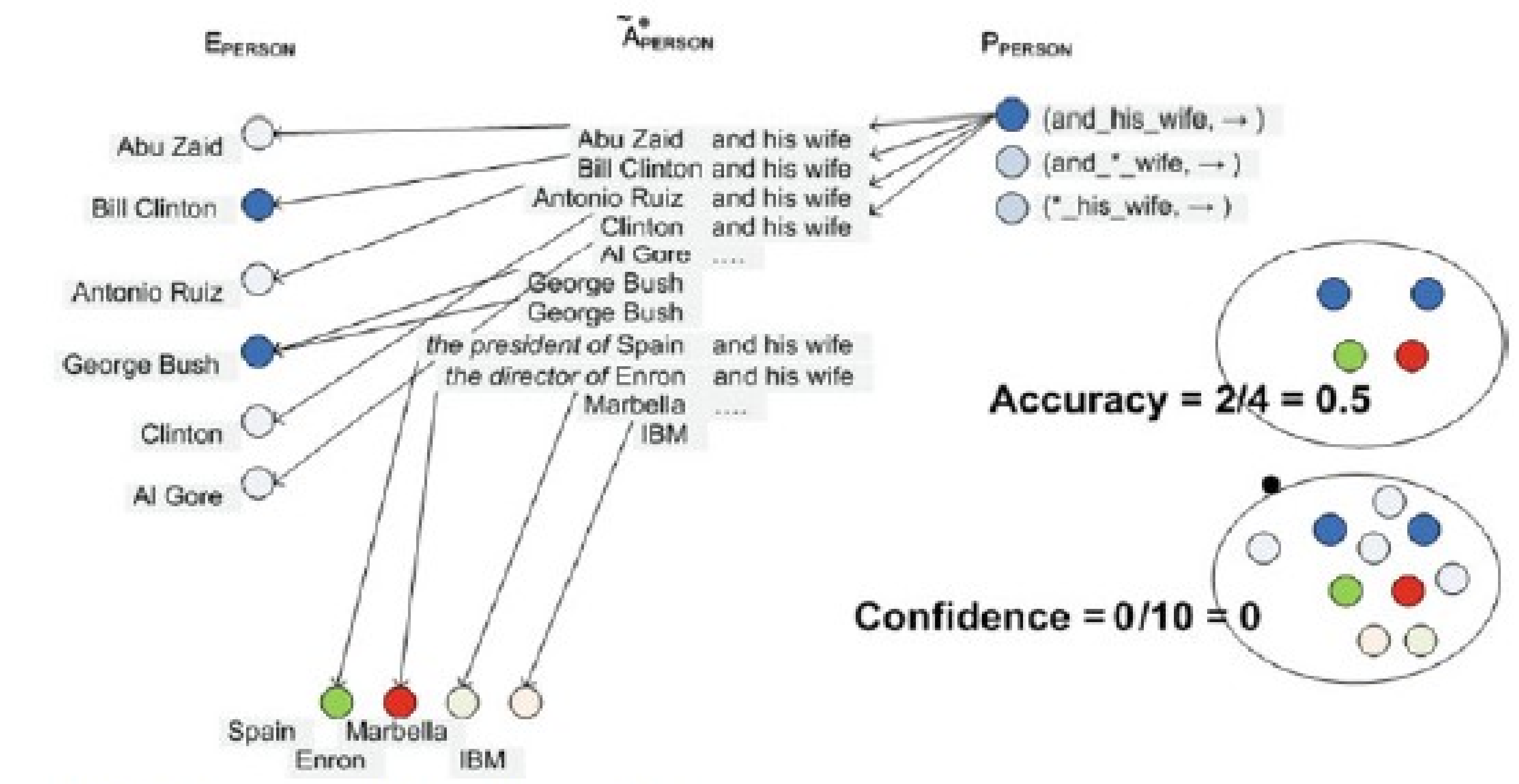


Fig. 2 The process of building the graph: Pattern expansion

2.7 Light Supervision

- (Sanchez et al. 2012, page 19)

Table 9 Name classification in CONLL-ES collection with Wiki dictionaries

	Baseline		Entities		Entities+Patterns	
	CONLL	ORG	PLO	PLOM	PLO	PLOM
P	26.27	–	**78.89**	77.82	73.42	73.86
R	56.48	–	47.34	46.64	**53.86**	53.75
F	35.86	–	59.17	58.33	62.14	**62.22**
Acc	–	39.34	61.30	61.01	62.60	**63.05**

Bold values indicate the best results in our experiments

3. Conclusion

- Progress has been made on
  - Synonym Detection
  - Seed evaluation
  - Rule evaluation
  - Resolving ambiguity
- Statistical models are still evolving
  - (Grave 2014) achieved F1 of 0.98, but only classification
- Active Learning increasingly relevant
- Deep Learning surprisingly absent
- Most of the presented techniques are available in OpenIE!

3. References

Collins, Michael and Singer, Yoram (1999). Unsupervised models for named entity classification. Proceedings of the joint SIGDAT conference on empirical methods in natural language processing and very large corpora, pages 100-110.

Curran, James R and Murphy, Tara and Scholz, Bernhard (2007). Minimising semantic drift with mutual exclusion bootstrapping. Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (3).

de Pablo-Sanchez, Cesar and Segura-Bedmar, Isabel and Martinez, Paloma and Iglesias-Maqueda, Ana (2013). Lightly supervised acquisition of named entities and linguistic patterns for multilingual text mining. Knowledge and information systems 35 (1), pages 87-109.

Grave, Edouard (2014). Weakly supervised named entity classification. Workshop on Automated Knowledge Base Construction (AKBC).

Jones, Rosie (2005). Learning to extract entities from labeled and unlabeled text. PhD Thesis, University of Utah.

3. References

Li, Xiaoli and Liu, Bing (2003). Learning to classify texts using positive and unlabeled data. IJCAI (3), pages 587-592.

Mintz, Mike and Bills, Steven and Snow, Rion and Jurafsky, Dan (2009). Distant supervision for relation extraction without labeled data. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (2), pages 1003-1011.

Nadeau, David and Sekine, Satoshi (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes 30 (1), pages 3-26.

Nigam, Kamal and McCallum, Andrew Kachites and Thrun, Sebastian and Mitchell, Tom (2000). Text classification from labeled and unlabeled documents using EM. Machine learning 39 (3), pages 103-134.

Riloff, Ellen and Jones, Rosie and others (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In proceedings of AAAI/IAAI, pages 474-479.