

Information Extraction

Лекция № *какой-то номер*

Elena Tutubalina

Kazan Federal University

Text classification at Different Granularities

- Text Categorization:
 - Classify an entire document
- Information Extraction (IE):
 - Identify and classify small units within documents
- Named Entity Extraction (NE):
 - A subset of IE
 - Identify and classify proper names
 - People, locations, organizations

What is Information Extraction

As a task:

Filling slots in a database from sub-segments of text.

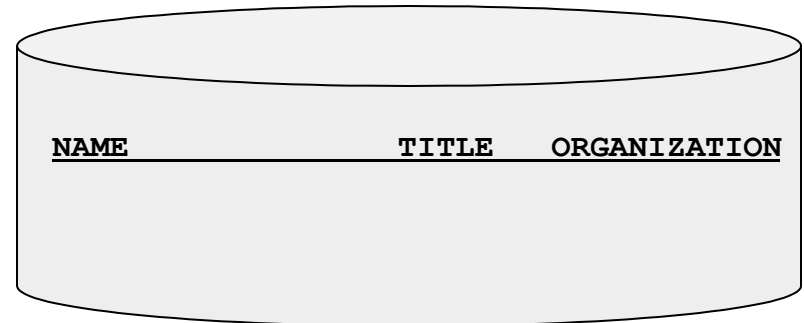
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



What is Information Extraction

As a task:

Filling slots in a database from sub-segments of text.

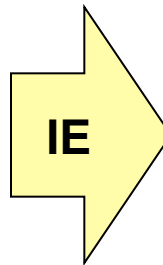
October 14, 2002, 4:00 a.m. PT

For years, [Microsoft Corporation](#) [CEO](#) [Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft](#) [VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...



<u>NAME</u>	<u>TITLE</u>	<u>ORGANIZATION</u>
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

What is Information Extraction

As a family
of techniques:

Information Extraction =
segmentation + classification + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...

Microsoft Corporation
CEO

Bill Gates

Microsoft

Gates

Microsoft

Bill Veghte

Microsoft

VP

Richard Stallman

founder

Free Software Foundation

aka "named entity
extraction"

What is Information Extraction

A family
of techniques:

Information Extraction =
segmentation + classification + association

October 14, 2002, 4:00 a.m. PT

For years, [Microsoft Corporation](#) [CEO](#) [Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, [Microsoft](#) claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. [Gates](#) himself says [Microsoft](#) will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft](#) [VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...

[Microsoft Corporation](#)
[CEO](#)

[Bill Gates](#)

[Microsoft](#)
[Gates](#)

[Microsoft](#)
[Bill Veghte](#)

[Microsoft](#)
[VP](#)

[Richard Stallman](#)
[founder](#)

[Free Software Foundation](#)

What is Information Extraction

A family
of techniques:

Information Extraction =
segmentation + classification + association

October 14, 2002, 4:00 a.m. PT

For years, [Microsoft Corporation](#) [CEO](#) [Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, [Microsoft](#) claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. [Gates](#) himself says [Microsoft](#) will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft](#) [VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...

[Microsoft Corporation](#)
[CEO](#)

[Bill Gates](#)

[Microsoft](#)
[Gates](#)

[Microsoft](#)

[Bill Veghte](#)

[Microsoft](#)
[VP](#)

[Richard Stallman](#)

[founder](#)

[Free Software Foundation](#)

INFORMATION EXTRACTION

- More general definition: extraction of structured information from unstructured documents
- IE Tasks:
 - Named entity extraction
 - Named entity recognition
 - Coreference resolution
 - Relationship extraction
- Semi-structured IE
 - Table extraction
- Terminology extraction











Landscape of IE Tasks:

Degree of Formatting

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Non-grammatical snippets, rich formatting & links

Barto, Andrew G.	(413) 545-2109	barto@cs.umass.edu	CS276
Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.			 
Berger, Emery D.	(413) 577-4211	emery@cs.umass.edu	CS344
Assistant Professor.			 
Brock, Oliver	(413) 577-0334	oli@cs.umass.edu	CS246
Assistant Professor.			 
Clarke, Lori A.	(413) 545-1328	clarke@cs.umass.edu	CS304
Professor. Software verification, testing, and analysis; software architecture and design.			 
Cohen, Paul R.	(413) 545-3638	cohen@cs.umass.edu	CS278
Professor. Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.			 

Grammatical sentences and some formatting & links

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

- Press
- **Contact**
- General information
- Directions maps

Tables

8:30 - 9:30 AM	Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty <i>Joseph Y. Halpern, Cornell University</i>				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
Cognitive Robotics	Logic Programming	Natural Language Generation	Complexity Analysis	Neural Networks	Games
739: A Logical Account of Causal and Topological Maps <i>Emilio Remolina and Benjamin Kuipers</i>	116: A-System: Problem Solving through Abduction <i>Marc Denecker, Antonis Kakas, and Bert Van Nuffelen</i>	758: Title Generation for Machine-Translated Documents <i>Rong Jin and Alexander G. Hauptmann</i>	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories <i>Marco Cadoli, Thomas Eiter, and Georg Gottlob</i>	179: Knowledge Extraction and Comparison from Local Function Networks <i>Kenneth McGarry, Stefan Wermter, and John MacIntyre</i>	71: Iterative Widening <i>Tristan Cazenave</i>
549: Online-Execution of ccGolog Plans <i>Henrik Grosskreutz and Gerhard Lakemeyer</i>	131: A Comparative Study of Logic Programs with Preference <i>Torsten Schaub and Kewen</i>	246: Dealing with Dependencies between Content Planning and Surface Realisation in a Pipeline Generation	470: A Perspective on Knowledge Compilation <i>Adnan Darwiche and Pierre Marquis</i>	258: Violation-Guided Learning for Constrained Formulations in Neural-Network Time-Series	353: Temporal Difference Learning Applied to a High Performance Game-Playing

Landscape of IE Tasks:

Intended Breadth of Coverage

Web site specific

Formatting

Amazon.com Book Pages

Genre specific

Layout

Resumes

Wide, non-specific

Language

University Names

The screenshot shows the Amazon.com product page for the book "Learning in Graphical Models" by Michael Irwin Jordan (Editor). The page features the Amazon logo, navigation tabs (WELCOME, YOUR STORE, BOOKS, ELECTRONICS, DVD, TOYS & GAMES), and a search bar. The book cover is displayed with a "LOOK INSIDE!" button. The price is listed as \$60.00, with a "NEW Super Saver Shipping FREE" banner. A "Great Buy" section at the bottom suggests buying the book with "Probabilistic Reasoning in Intelligent Systems" for a total price of \$128.95.

The screenshot displays two resumes side-by-side. The top resume is for Jason D. M. Rennie, showing his contact information, research interests in data analysis and classification, and his current position at MIT AI Lab. The bottom resume is for L. Douglas Baker, detailing his contact information, education at Carnegie Mellon University and the Technical University of Berlin, and his research experience in machine learning and information retrieval.

The screenshot shows a conference schedule for a day with sessions from 8:30 AM to 11:30 AM. The sessions include an invited talk by Joseph Y. Halpern, a coffee break, and technical paper sessions on Cognitive Robotics, Logic Programming, Natural Language Generation, and Complexity Analysis. Below the schedule, there is a contact section for Dr. Steven Minton, Founder/CTO, and Frank Huybrechts, COO, providing their roles and affiliations.

Landscape of IE Tasks”

Complexity

Closed set

U.S. states

He was born in Alabama...

The big Wyoming sky...

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

...was among the six houses
sold by Hope Feldman that year.

Pawel Opalinski, Software
Engineer at WhizBang Labs.

Landscape of IE Tasks:

Single Field/Record

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity

Person: Jack Welch

Person: Jeffrey Immelt

Location: Connecticut

Binary relationship

Relation: Person-Title

Person: Jack Welch

Title: CEO

Relation: Company-Location

Company: General Electric

Location: Connecticut

N-ary record

Relation: Succession

Company: General Electric

Title: CEO

Out: Jack Welsh

In: Jeffrey Immelt

“Named entity” extraction

Adapted from slide by William
Cohen

State of the Art Performance: a sample

- Named entity recognition from newswire text
 - Person, Location, Organization, ...
 - F1 in high 80' s or low- to mid-90' s
- Binary relation extraction
 - Contained-in (Location1, Location2)
Member-of (Person1, Organization1)
 - F1 in 60' s or 70' s or 80' s
- Web site structure recognition
 - Extremely accurate performance obtainable
 - Human effort (~10min?) required on each site

Three generations of IE systems

- Hand-Built Systems – Knowledge Engineering [1980s–]
 - Rules written by hand
 - Require experts who understand both the systems and the domain
 - Iterative guess-test-tweak-repeat cycle
- Automatic, Trainable Rule-Extraction Systems [1990s–]
 - Rules discovered automatically using predefined templates, using automated rule learners
 - Require huge, labeled corpora (effort is just moved!)
- Statistical Models [1997 –]
 - Use machine learning to learn which features indicate boundaries and types of entities.
 - Learning usually supervised; may be partially unsupervised

Named Entity Recognition (NER)

A **named entity** is a word or a word collocation that means a specific object or an event and distinguishes it from other similar objects.

1. Президент [Владимир Путин] PER 17 декабря провел традиционную пресс-конференцию перед Новым Годом.
2. Студенты и Татьяны получают эксклюзивный пропуск на Главный каток страны.

Input:

Apple Inc., formerly Apple Computer, Inc., is an American multinational corporation headquartered in Cupertino, California that designs, develops, and sells consumer electronics, computer software and personal computers. It was established on April 1, 1976, by Steve Jobs, Steve Wozniak and Ronald Wayne.

Output:

Apple Inc., formerly Apple Computer, Inc., is an American multinational corporation headquartered in Cupertino, California that designs, develops, and sells consumer electronics, computer software and personal computers. It was established on April 1, 1976, by Steve Jobs, Steve Wozniak and Ronald Wayne.

Named Entity Recognition (NER)

- Locate and classify atomic elements in text into predefined categories (persons, organizations, locations, temporal expressions, quantities, percentages, monetary values, ...)
- Input: a block of text
 - *Jim bought 300 shares of Acme Corp. in 2006.*
- Output: annotated block of text
 - `<ENAMEX TYPE="PERSON">Jim</ENAMEX> bought <NUMEX TYPE="QUANTITY">300</NUMEX> shares of <ENAMEX TYPE="ORGANIZATION">Acme Corp.</ENAMEX> in <TIMEX TYPE="DATE">2006</TIMEX>`
 - ENAMEX tags (MUC in the 1990s)

HOW

- Two tasks:
 - Identifying the part of text that mentions a text (RECOGNITION)
 - Classifying it (CLASSIFICATION)
- The two tasks are reduced to a standard classification task by having the system classify WORDS

Basic Problems in NER

- Variation of NEs – e.g. John Smith, Mr Smith, John.
- Ambiguity of NE types
 - John Smith (company vs. person)
 - May (person vs. month)
 - Washington (person vs. location)
 - 1945 (date vs. time)
- Ambiguity with common words, e.g. “may”

Problems in NER

- Category definitions are intuitively quite clear, but there are many grey areas.
- Many of these grey area are caused by **metonymy**.

Organisation vs. Location : “**England** won the World Cup” vs. “The World Cup took place in **England**”.

Company vs. Artefact: “shares in **MTV**” vs. “watching **MTV**”

Location vs. Organisation: “she met him at **Heathrow**” vs. “the **Heathrow** authorities”

More complex problems in NER

- Issues of style, structure, domain, genre etc.
 - Punctuation, spelling, spacing, formatting,all have an impact

Dept. of Computing and Maths
Manchester Metropolitan University
Manchester
United Kingdom

> Tell me more about Leonardo
> Da Vinci

Approaches to NER:

List Lookup

- System that recognises only entities stored in its lists (GAZETTEERS).
- Advantages - Simple, fast, language independent, easy to retarget
- Disadvantages – collection and maintenance of lists, cannot deal with name variants, cannot resolve ambiguity

GAZETTEERS

Vocabulary	Size, objects	Clarification	Examples
Famous persons	31482	Famous people	Владимир Путин
First names	2773	First names	Василий, Анна, Том
Surnames	66108	Surnames	Кузнецов, Грибоедов
Verbs of informing	1729	Verbs that usually occur with persons	высказать, признаться
Companies	33380	Organization names	Сбербанк
Company types	6774	Organization types	организация, авиафирма
Geography	8969	Geographical objects	Балтийское море
Equipment	44094	Devices, equipment, tools	устройство, телефон

Approaches to NER:

Shallow Parsing

- Names often have internal structure. These components can be either stored or guessed.

location:

CapWord + {City, Forest, Center}

e.g. Sherwood Forest

Cap Word + {Street, Boulevard, Avenue, Crescent, Road}

e.g. Portobello Street

Shallow Parsing Approach

(E.g., Mikheev et al 1998)

- External evidence - names are often used in very predictive local contexts

Location:

“to the” COMPASS “of” CapWord

e.g. *to the south of **Loitokitok***

“based in” CapWord

e.g. *based in **Loitokitok***

CapWord “is a” (ADJ)? GeoWord

e.g. ***Loitokitok** is a friendly city*

Difficulties in Shallow Parsing

Approach

- **Ambiguously capitalised words** (first word in sentence)

[All American Bank] vs. All [State Police]

- **Semantic ambiguity**

“John F. Kennedy” = airport (location)

“Philip Morris” = organisation

- **Structural ambiguity**

[Cable and Wireless] vs. [Microsoft] and [Dell]

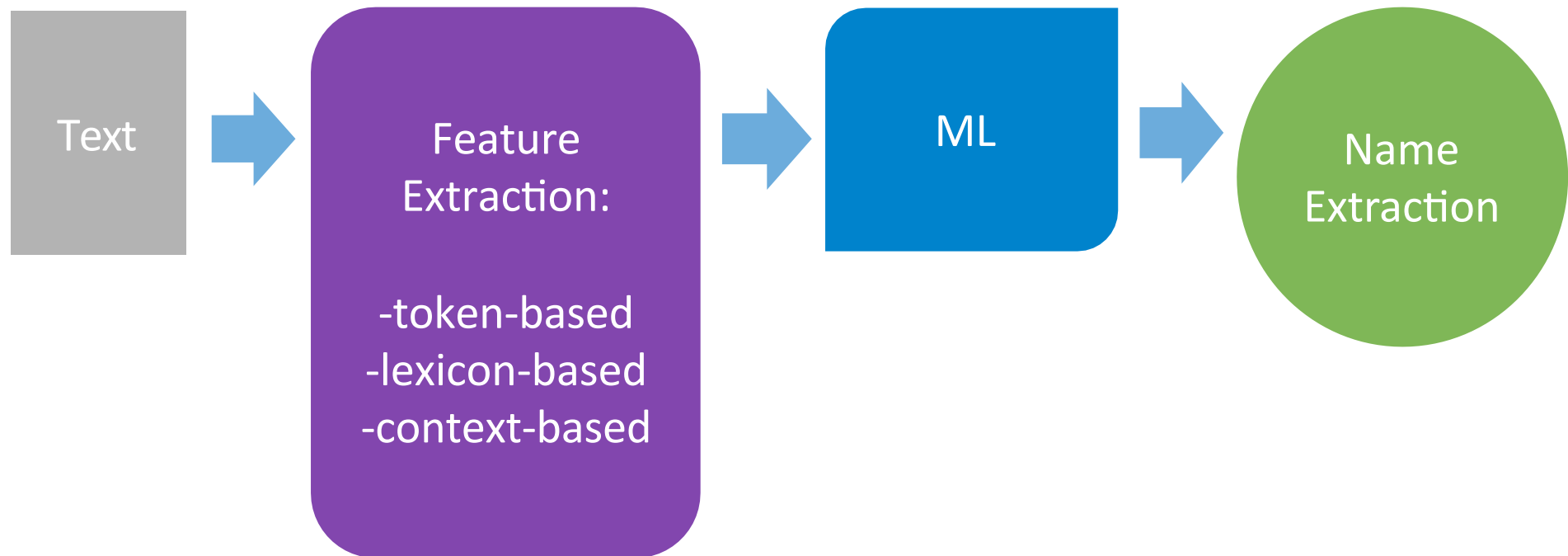
[Center for Computational Linguistics] vs. message
from [City Hospital] for

[John Smith].

Machine learning approaches to NER

- NER as classification: the IOB representation
- Supervised methods
 - Support Vector Machines
 - Logistic regression (aka Maximum Entropy)
 - Sequence pattern learning
 - Hidden Markov Models
 - Conditional Random Fields
- Distant learning
- Semi-supervised methods

Scheme of text processing



Labeling representation

IO-scheme (Inside-Outside)

- **I** - belongs to named entity
- **O** - does not belong to named entity

$|C| + 1$ classes

BIO-scheme (Begin-Inside-Outside)

- **B** - named entity beginning
- **I** - named entity continuation
- **O** - not named entity

$2 * |C| + 1$ classes

Token	IO-Labels	BIO-labels
Владимир	I-PER	B-PER
Путин	I-PER	I-PER
посетил	O-OUTSIDE	O-OUTSIDE
Англию	I-GEOPOLIT	B-GEOPOLIT

FEATURES

Most traditional features

1. Token initial form (lemma)
2. Number of symbols in a token
3. Letter case: BigBig, BigSmall, SmallSmall, Fence
4. Token type
 - part of speech
 - type of punctuation
5. The presence of a vowel (a binary feature)
6. If a token contains a known letter n-gram from a pre-defined set:
 - Кузнец^{ов}, Матви^{енко}, Джуга^{швили}
 - ^{Го}с^{де}партамент, Газ^{про}м

FEATURES

For each running word:

- **WORD**: the word itself (both unchanged and lower-cased)
e.g. Casa casa
- **POS**: the part of speech of the word (as produced by TagPro)
e.g. Oggi SS (singular noun)
- **AFFIX**: prefixes/suffixes (1, 2, 3 or 4 chars. at the start/end of the word)
e.g. Oggi {o,og,ogg,oggi, – i,gi,ggi,oggi}
- **ORTHOgraphic** information (e.g. capitalization, hyphenation)
e.g. Oggi C (capitalized)
oggi L (lowercased)

FEATURES

- **COLLOC**ation bigrams
 - 36.000, Italian newspapers ranked by MI values
- **Gazz**etters
 - **PERSONS**: Person proper names or titles
(154.000, Italian phone-book, Wikipedia,)
 - **TOWNS**: World (main), Italian (comuni) and Trentino's (frazioni) towns (12.000, from various internet sites)
 - **STOCK-MARKET**: Italian and American stock market organizations (5.000, from stock market sites)
 - **WIKI-GEO**: Wikipedia geographical locations (3.200,)

Context features and example

Token	Lemma	Register	Token Type	Second Name	Geo	Label
В	В	Small	Auxiliary	False	False	NO
России	РОССИЯ	BigSmall	Noun	False	Geo1	GEOPOLIT
Алиев	АЛИЕВ	BigSmall	Noun	Sname1	False	PER
третий	ТРЕТИЙ	Small	Numeral	False	False	NO
раз	РАЗ	Small	Auxiliary	False	False	NO

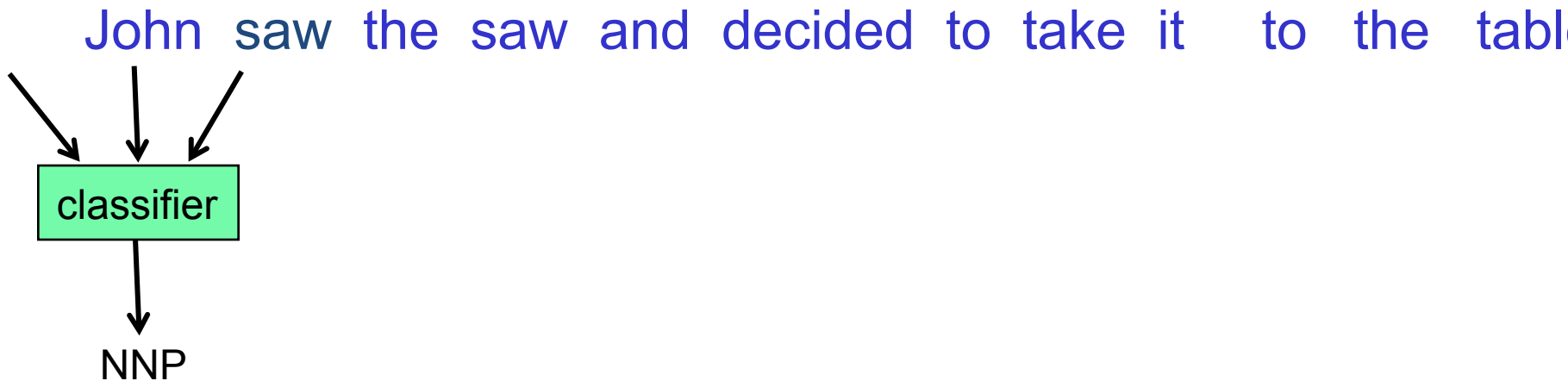
Supervised ML for NER

- Methods already seen
 - Decision trees
 - Support Vector Machines
- **Sequence learning**
 - Hidden Markov Models
 - Maximum Entropy Models
 - **Conditional Random Fields**

NER as a SEQUENCE CLASSIFICATION TASK

Sequence Labeling as Classification: POS Tagging

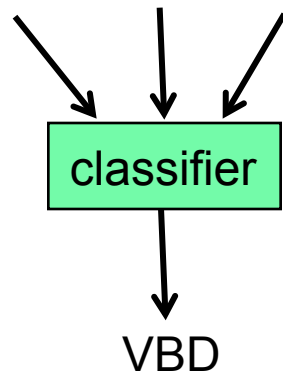
- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).



Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

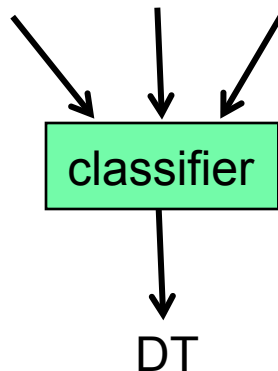
John saw the saw and decided to take it to the table



Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

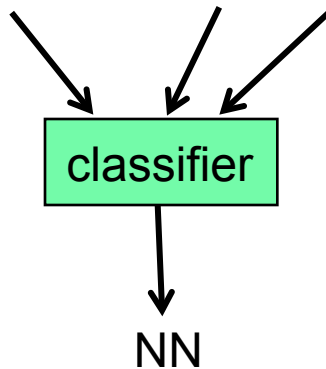
John saw the saw and decided to take it to the table



Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

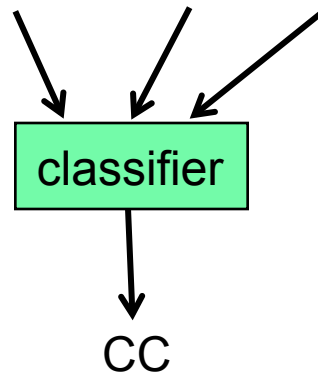
John saw the saw and decided to take it to the table



Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

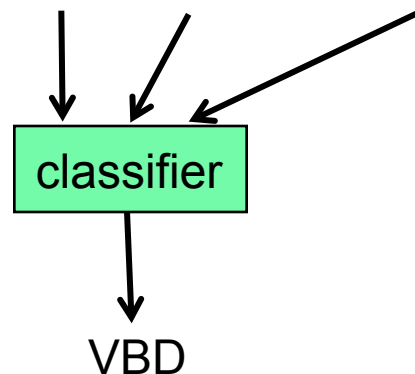
John saw the saw and decided to take it to the table



Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

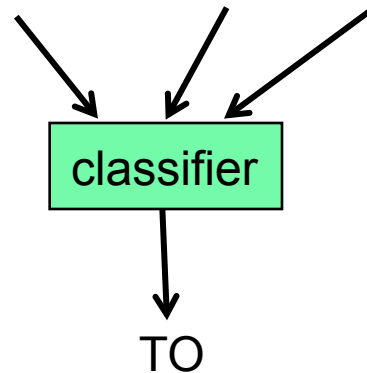
John saw the saw and decided to take it to the table



Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

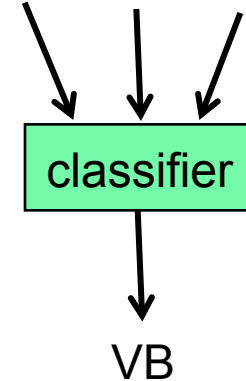
John saw the saw and decided to take it to the table



Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

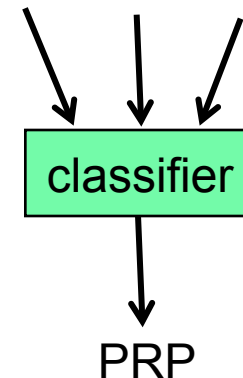
John saw the saw and decided to take it to the table



Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

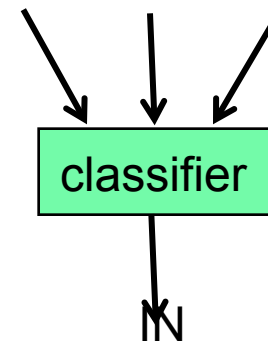
John saw the saw and decided to take it to the table.



Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

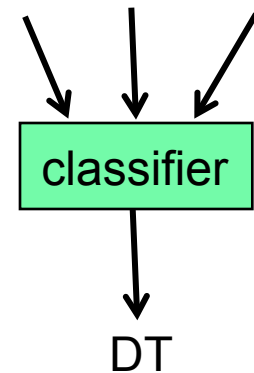
John saw the saw and decided to take it to the table



Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

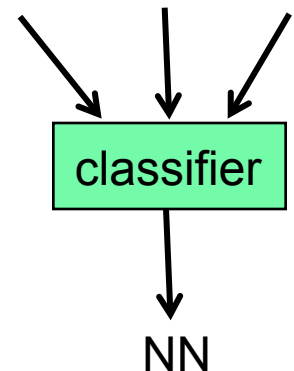
John saw the saw and decided to take it to the table



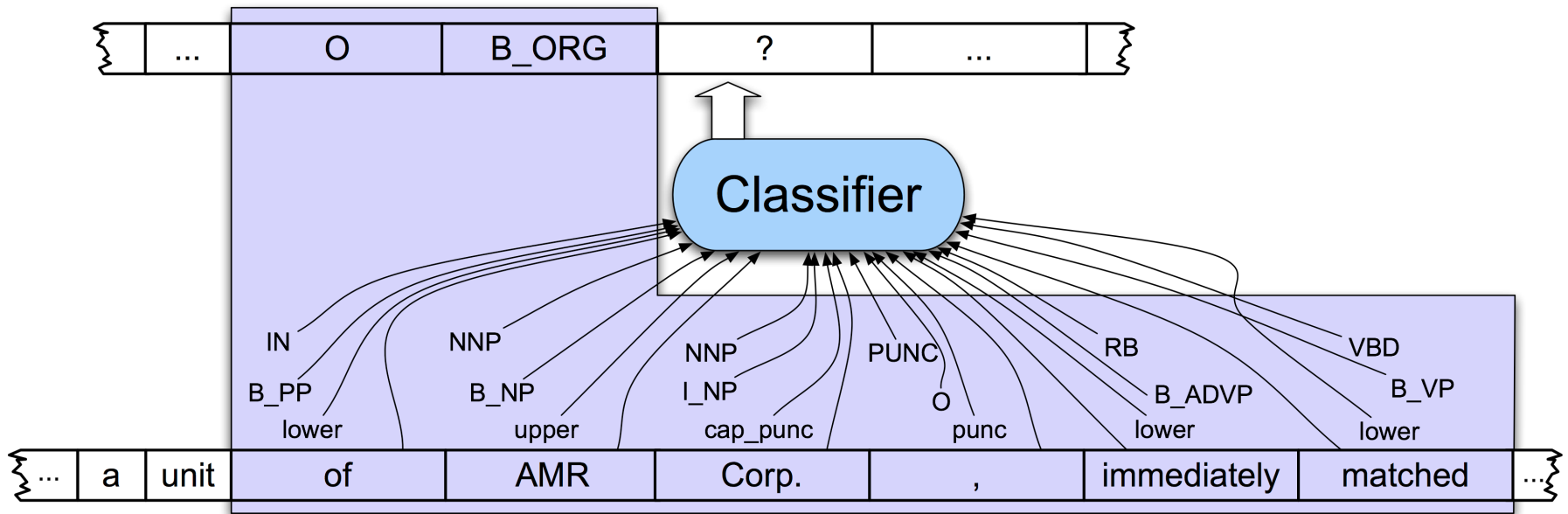
Sequence Labeling as Classification

- Classify each token independently but use as input features, information about the surrounding tokens (sliding window).

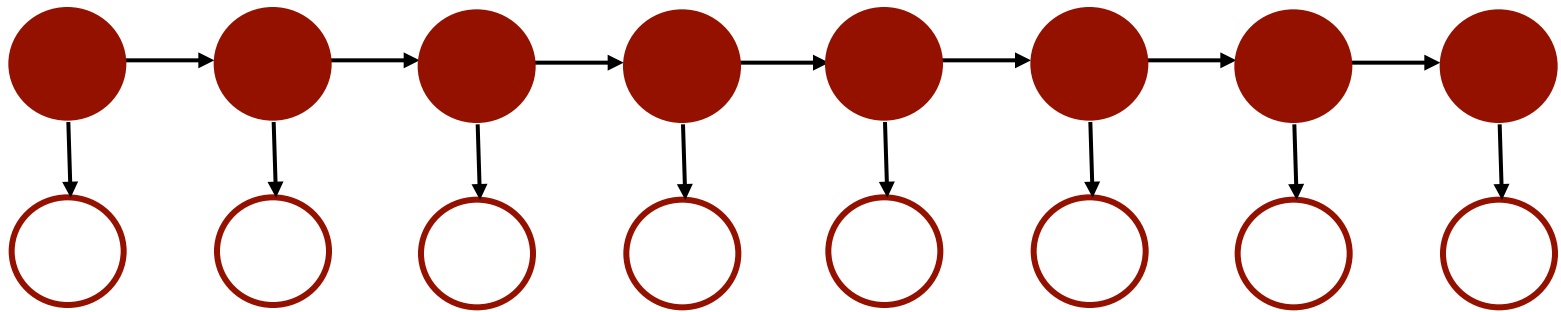
John saw the saw and decided to take it to the table



NER as Sequence Labeling

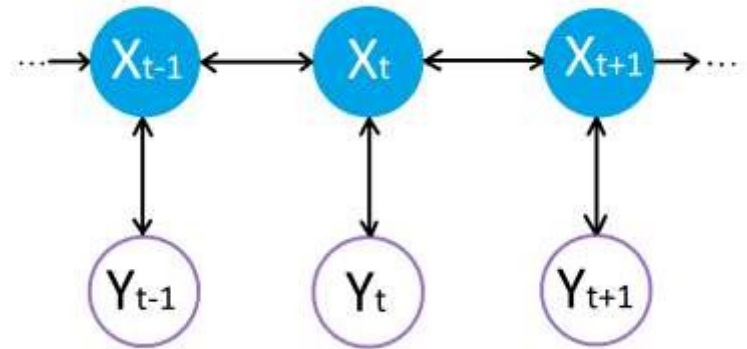
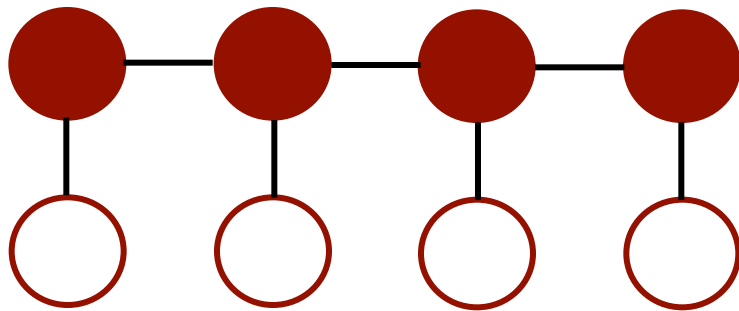


Hidden Markov Models (HMMs)



- Generative
 - Find parameters to maximize $P(X,Y)$
- Assumes features are independent
- When labeling X_i future observations are taken into account (forward-backward)

Conditional Random Fields (CRFs)



- Discriminative
 - Find parameters to maximize $P(Y|X)$
 - Doesn't assume that features are independent
 - When labeling Y_i future observations are taken into account
- ➔ The best of both worlds!

Discriminative Vs. Generative

$p(\mathbf{y}, \mathbf{x})$

- **Generative Model:** A model that generate observed data randomly
- **Naïve Bayes:** once the class label is known, all the features are independent

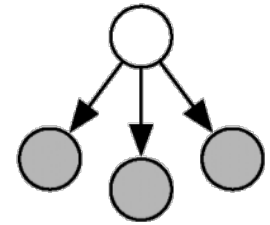
$$p(y, \mathbf{x}) = p(y) \prod_{k=1}^K p(x_k | y)$$

- **Discriminative:** Directly estimate the posterior probability; Aim at modeling the “discrimination” between different outputs

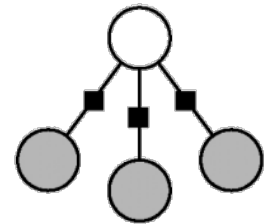
$p(\mathbf{y} | \mathbf{x})$

- **MaxEnt** classifier: linear combination of feature function in the exponent,

$$p(y | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left\{ \sum_{k=1}^K \theta_k f_k(y, \mathbf{x}) \right\}$$



Naive Bayes

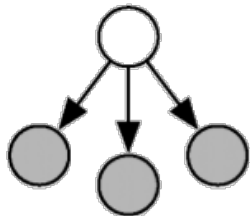


Logistic Regression

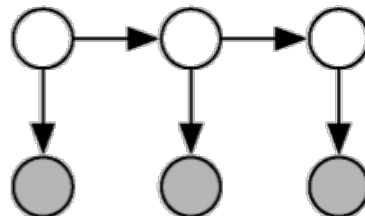
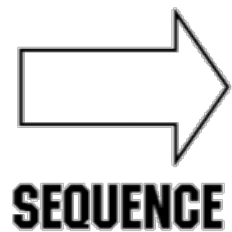
Both generative models and discriminative models describe distributions over (y, \mathbf{x}) , but they work in different directions.

Discriminative Vs. Generative

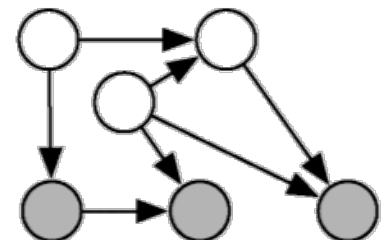
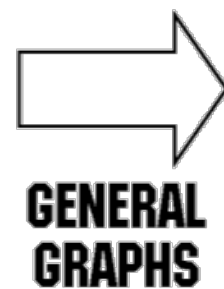
$p(\mathbf{y}, \mathbf{x})$



Naive Bayes



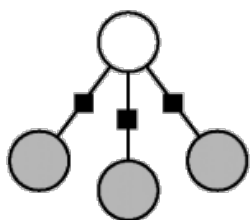
HMMs



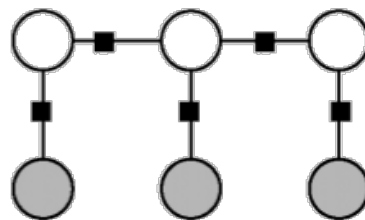
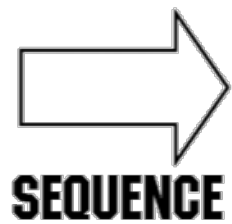
Generative directed model:



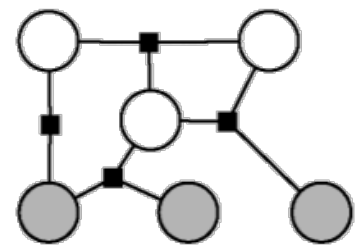
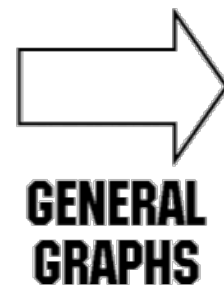
$p(\mathbf{y} | \mathbf{x})$



logistic Regression



Linear-chain CRFs



General CRFs

○=observable

○=unobservable

Target metric

$$Precision = \frac{intersectionCount}{classifierCount}$$

$$Recall = \frac{intersectionCount}{expertCount}$$

$$Fscore = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

intersectionCount is the number of named entities labeled by both:
the classifier and the expert;

classifierCount is the number of named entities labeled by only the
classifier;

expertCount is the number of named entities labeled by only the
expert.

Text collections

- "Persons-1000" (1000 news documents)
 - Russian names: Александр Игнатенко, Алексей Волков
- "Persons-1111F" (1111 news documents)
 - Eastern names: Абдалла Халаф, Иттё Ито

We additionally labeled:

- Organizations (ORG)
- Media organizations having a specific function of information providing (MEDIA)
- Locations (LOC)
- States and capitals in the role of a state (GEOPOLIT)

Experiments on Collection “Persons-1000”

NE Type	F-score, %		
	IO	IO + rules	BIO
PER	94.95	95.09	96.08
ORG	80.03	80.23	83.84
LOC	92.60	92.60	94.57
Average	89.54	89.67	91.71

NE Type	F-score, %		
	IO	IO + rules	BIO
PER	94.95	95.01	95.63
ORG	75.90	76.16	80.06
MEDIA	87.95	87.95	87.99
LOC	84.53	84.53	86.91
GEOPOLIT	94.65	94.65	94.50
Average	88.21	88.37	89.93

Cross-validation
3:1

Experiments on collection with Eastern names (Persons-1111F)

Person name extraction

Collection	F-score, %	
	Rule-based (Trofimov, 2014)	Our system
Pesons-1000	96.62	96.08
Persons-1111F	64.43	81.68

“Persons-1000”: cross-validation 3:1

“Persons-1111F” : training on “Persons-1000”

NEW DOMAINS

- BIOMEDICAL
- CHEMISTRY
- HUMANITIES: MORE FINE GRAINED TYPES

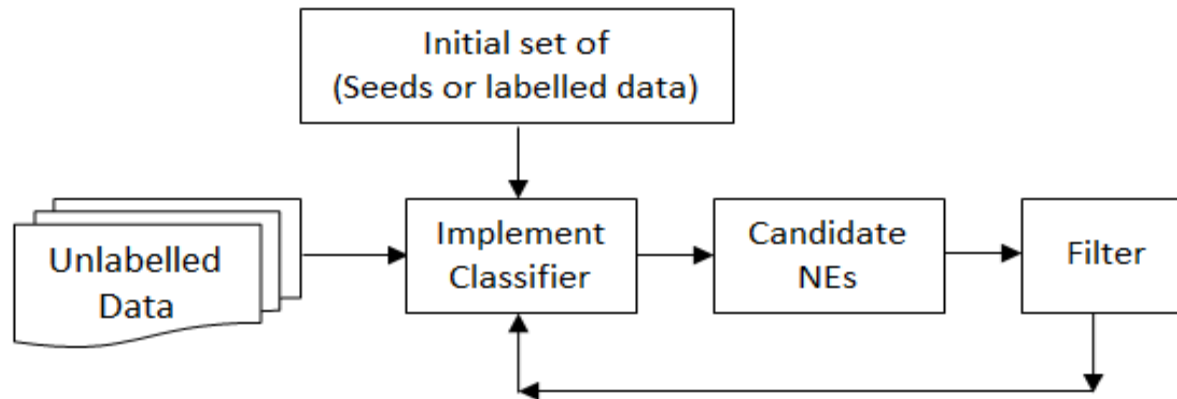
Bioinformatics Named Entities

- Protein
- DNA
- RNA
- Cell line
- Cell type
- Drug
- Chemical

Semi-supervised learning

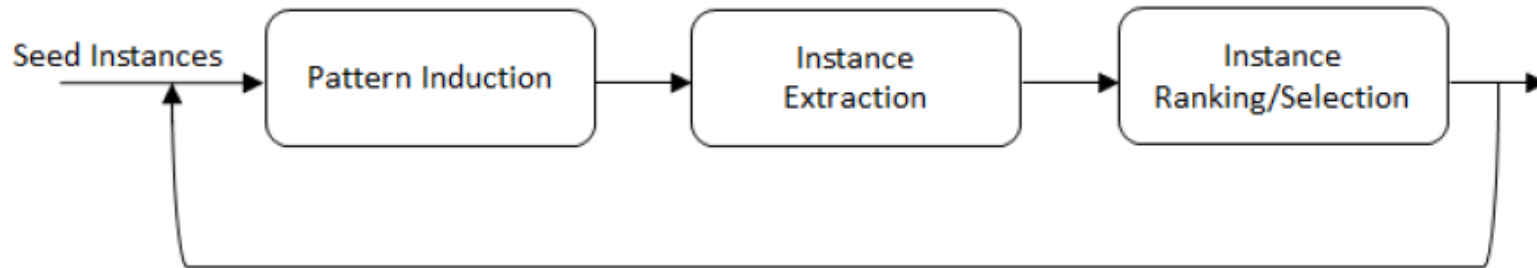
- Modest amounts of supervision
 - Small size of training data
 - Supervisor input sought when necessary
- Aims to match supervised learning performance, but with much less human effort
- Bootstrapping
 - Seeds used to identify contextual clues
 - Contextual clues used to find more NEs

Semi-supervised learning



- **Examples:** (Brin 1998); (Collins and Singer 1999); (Riloff and Jones 1999); (Cucchiarelli and Velardi 2001); (Pasca *et al.* 2006); (Heng and Grishman 2006); (Nadeau *et al.* 2006), and (Liao and Veeramachaneni, 2009)

ASemiNER - Methodology



Input

—A seed list of a few examples of a given NE type

- ‘Muhammad’ & ‘Obama’ can be used as seed examples for entity of type person.

Parameters

—Number of iterations!

—Number of initial seeds!

—The ranking measure (Reliability measure)!