Appendix 2 to lecture 18 of “Machine Learning”

I256:
Applied Natural Language Processing

Marti Hearst
Nov 15, 2006

Today

- Information Extraction
  - What it is
  - Historical roots: MUC
  - Current state-of-art performance
  - Various Techniques

Classifying at Different Granularies

- 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.**

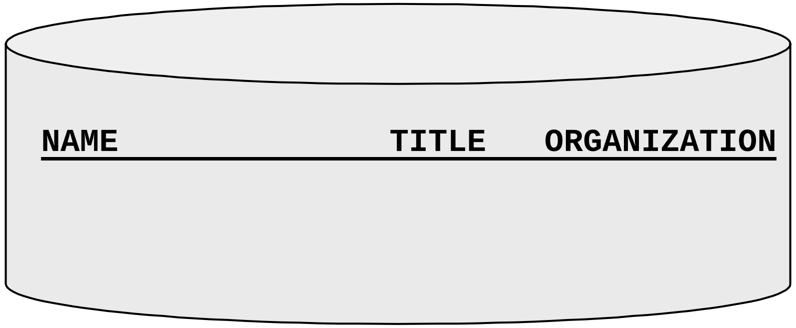
October 14, 2002, 4:00 a.m. PT

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…

IE

NAME	TITLE	ORGANIZATION
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.

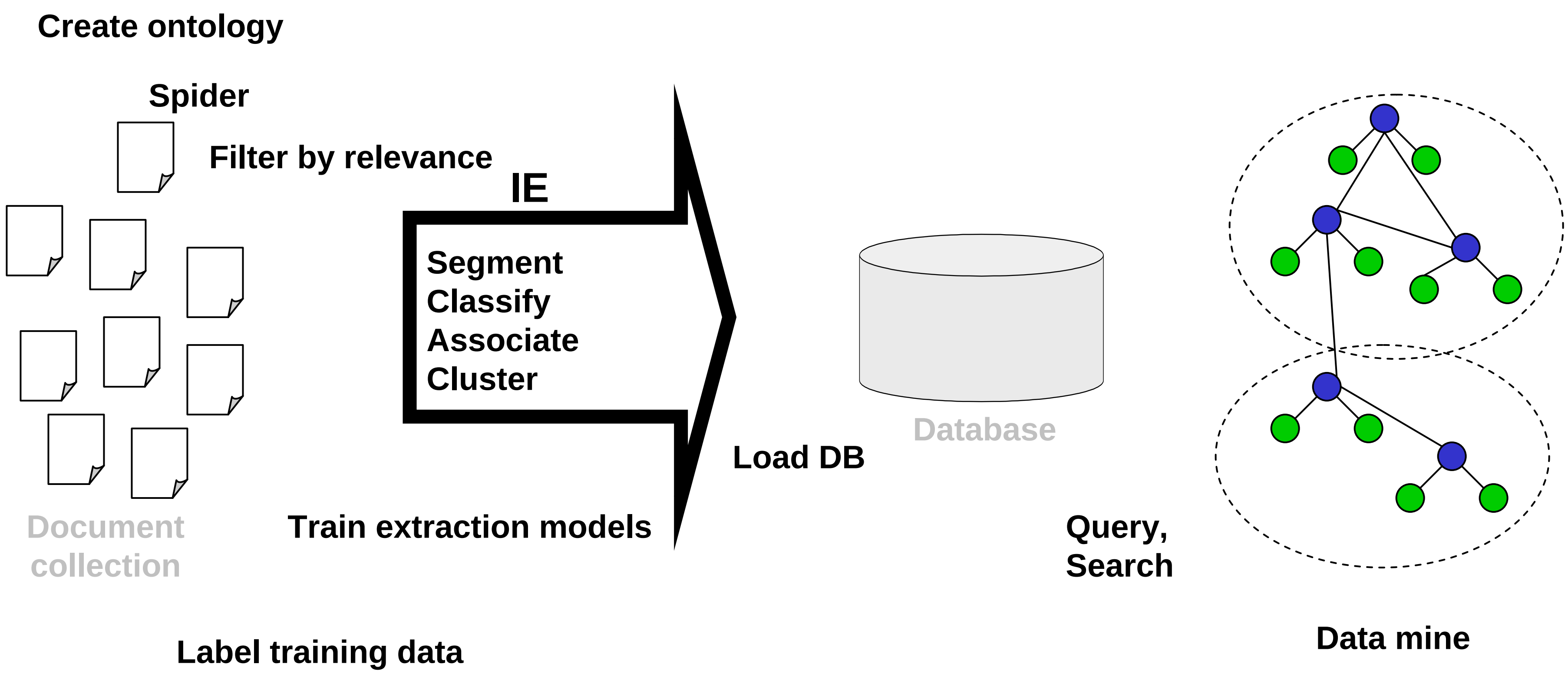
Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"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

IE in Context



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.

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

Non-grammatical snippets, rich formatting & links

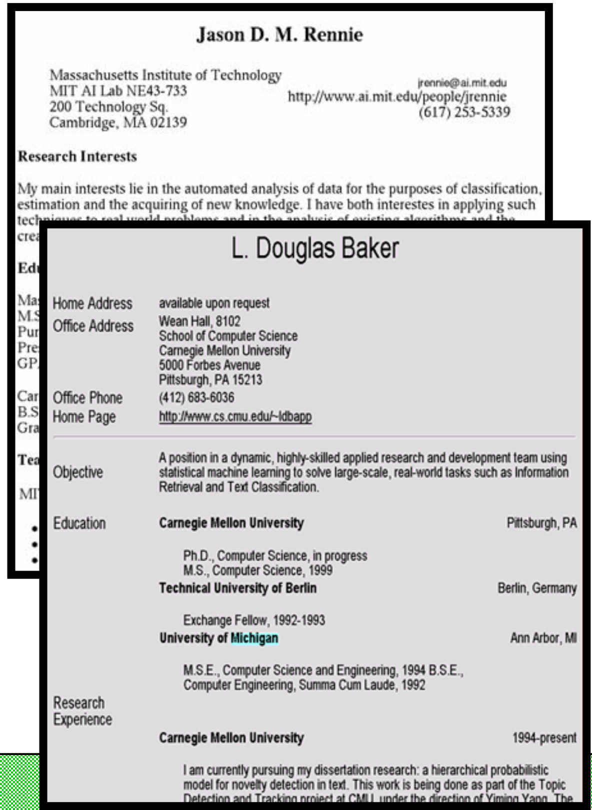
Barto, Andrew G.	(413) 545-2109	barto@cs.umass.edu	CS276
Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.			
Berger, Emery D.	(413) 577-4211	emery@cs.umass.edu	CS344
Assistant Professor.			
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.			

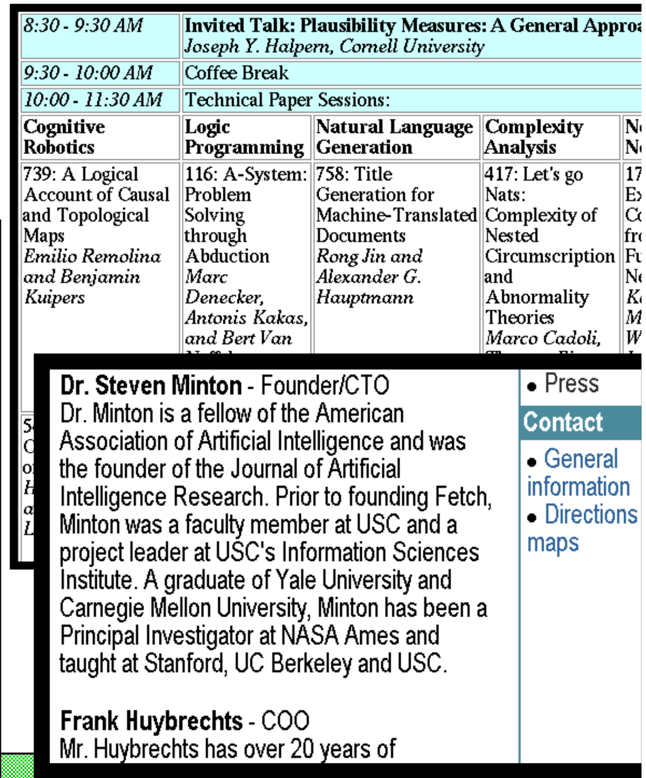
Tables

8:30 - 9:30 AM	**Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty** *Joseph Y. Halpern, Cornell University*				
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 *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas, and Bert Van Nuffelen*	758: Title Generation for Machine-Translated Documents *Rong Jin and Alexander G. Hauptmann*	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories *Marco Cadoli, Thomas Eiter, and Georg Gottlob*	179: Knowledge Extraction and Comparison from Local Function Networks *Kenneth McGarry, Stefan Wermter, and John MacIntyre*	71: Iterative Widening *Tristan Cazenave*
549: Online-Execution of ccGolog Plans *Henrik Grosskreutz*	131: A Comparative Study of Logic Programs with	246: Dealing with Dependencies between Content Planning and	470: A Perspective on Knowledge Compilation	258: Violation-Guided Learning for Constrained	353: Temporal Difference Learning Applied to a

Landscape of IE Tasks: Intended Breadth of Coverage

Web site specific	**Genre specific**	**Wide, non-specific**
Formatting	Layout	Language
Amazon.com Book Pages	Resumes	University Names





Adapted from slide by William Cohen

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

MUC: the genesis of IE

- DARPA funded significant efforts in IE in the early to mid 1990's.
- Message Understanding Conference (MUC) was an annual event/competition where results were presented.
- Focused on extracting information from news articles:
  - Terrorist events
  - Industrial joint ventures
  - Company management changes
- Information extraction of particular interest to the intelligence community (CIA, NSA). (Note: early '90's)

Message Understanding Conference (MUC)

- Named entity
  - Person, Organization, Location
- Co-reference
  - Clinton ↔ President Bill Clinton
- Template element
  - Perpetrator, Target
- Template relation
  - Incident
- Multilingual

MUC Typical Text

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan. The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production of 20,000 iron and "metal wood" clubs a month

MUC Typical Text

Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan. The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production of 20,000 iron and "metal wood" clubs a month

MUC Templates

- Relationship
  – tie-up
- Entities:
  – Bridgestone Sports Co, a local concern, a Japanese trading house
- Joint venture company
  – Bridgestone Sports Taiwan Co
- Activity
  – ACTIVITY 1
- Amount
  – NT$2,000,000

MUC Templates

- ATIVITY 1
  - Activity
    – Production
  - Company
    – Bridgestone Sports Taiwan Co
  - Product
    – Iron and “metal wood” clubs
  - Start Date
    – January 1990

Example of IE from FASTUS (1993)

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

Example of IE: FASTUS(1993)

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Example of IE: FASTUS(1993): Resolving anaphora

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

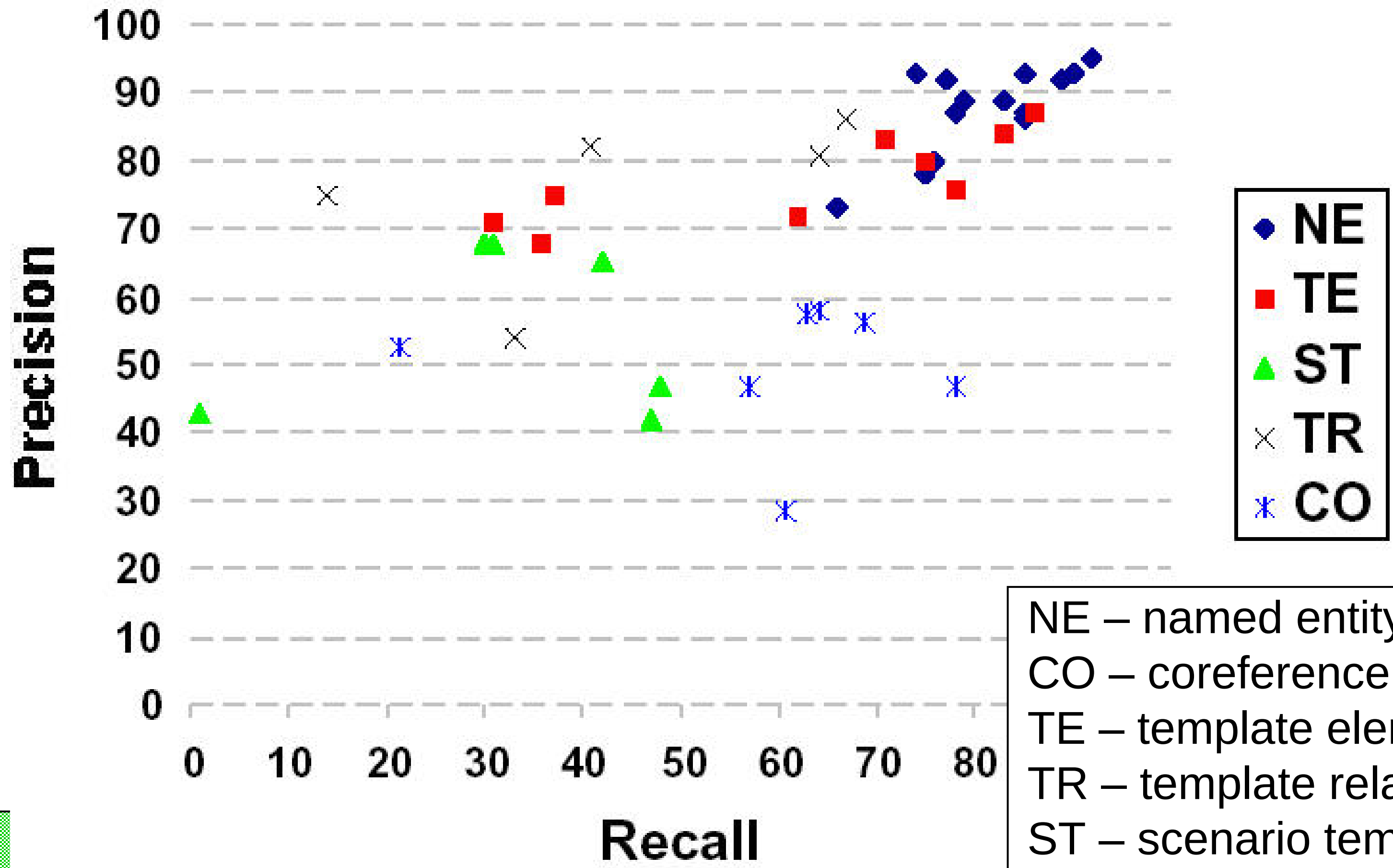
TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Evaluating IE Accuracy

- Always evaluate performance on independent, manually-annotated test data not used during system development.
- Measure for each test document:
  - Total number of correct extractions in the solution template: N
  - Total number of slot/value pairs extracted by the system: E
  - Number of extracted slot/value pairs that are correct (i.e. in the solution template): C
- Compute average value of metrics adapted from IR:
  - Recall = C/N
  - Precision = C/E
  - F-Measure = Harmonic mean of recall and precision

MUC Information Extraction: State of the Art c. 1997



NE – named entity recognition
CO – coreference resolution
TE – template element construction
TR – template relation construction
ST – scenario template production

Two kinds of NE approaches

Knowledge Engineering

- rule based
- developed by experienced language engineers
- make use of human intuition
- requires only small amount of training data
- development could be very time consuming
- some changes may be hard to accommodate

Learning Systems

- use statistics or other machine learning
- developers do not need LE expertise
- requires large amounts of annotated training data
- some changes may require re-annotation of the entire training corpus
- annotators are cheap (but you get what you pay for!)

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

Trainable IE systems

Pros

- Annotating text is simpler & faster than writing rules.
- Domain independent
- Domain experts don't need to be linguists or programers.
- Learning algorithms ensure full coverage of examples.

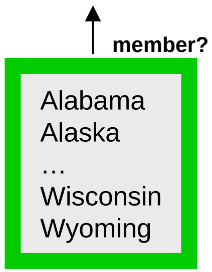
Cons

- Hand-crafted systems perform better, especially at hard tasks (but this is changing).
- Training data might be expensive to acquire.
- May need huge amount of training data.
- Hand-writing rules isn't ***that*** hard!!

Landscape of IE Techniques

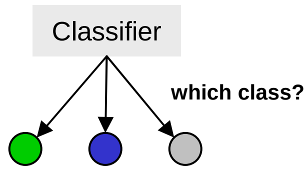
Lexicons

Abraham Lincoln was born in Kentucky.



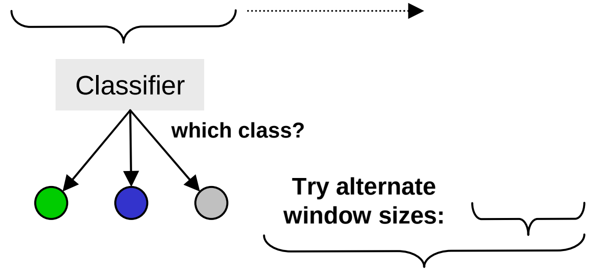
Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.

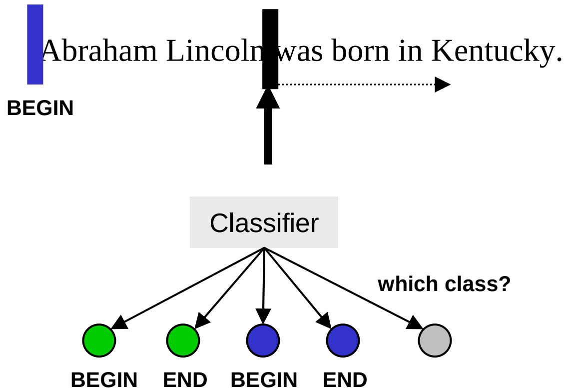


Sliding Window

Abraham Lincoln was born in Kentucky.

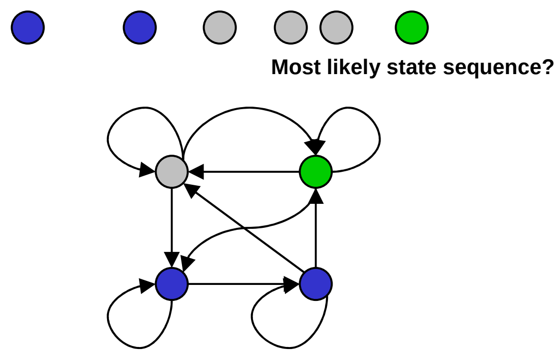


Boundary Models



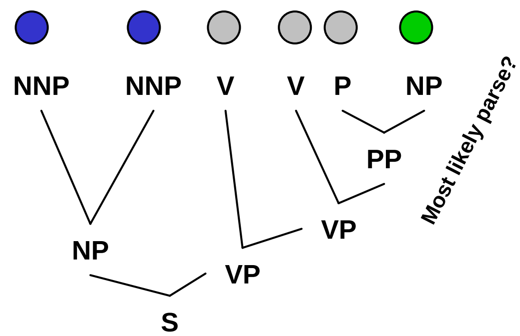
Finite State Machines

Abraham Lincoln was born in Kentucky.



Context Free Grammars

Abraham Lincoln was born in Kentucky.



Any of these models can be used to capture words, formatting or both.

Adapted from slide by William Cohen

Successors to MUC

- CoNNL: Conference on Computational Natural Language Learning
  - Different topics each year
  - 2002, 2003: Language-independent NER
  - 2004: Semantic Role recognition
  - 2001: Identify clauses in text
  - 2000: Chunking boundaries
    - http://cnts.uia.ac.be/conll2003/ (also conll2004, conll2002…)
    - Sponsored by SIGNLL, the Special Interest Group on Natural Language Learning of the Association for Computational Linguistics.
- ACE: *Automated Content Extraction*
  - Entity Detection and Tracking
    - Sponsored by NIST
    - http://wave.ldc.upenn.edu/Projects/ACE/
- Several others recently
  - See http://cnts.uia.ac.be/conll2003/ner/

State of the Art Performance: examples

- 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

CoNNL-2003

- Goal: identify boundaries and types of named entities
  - People, Organizations, Locations, Misc.

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] .

  - Experiment with incorporating external resources (Gazeteers) and unlabeled data
- Data:
  - Using IOB notation
  - 4 pieces of info for each term

Word	POS	Chunk	EntityType
U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

Details on Training/Test Sets

English data	Articles	Sentences	Tokens
Training set	946	14,987	203,621
Development set	216	3,466	51,362
Test set	231	3,684	46,435

German data	Articles	Sentences	Tokens
Training set	553	12,705	206,931
Development set	201	3,068	51,444
Test set	155	3,160	51,943

Table 1: Number of articles, sentences and tokens in each data file.

English data	LOC	MISC	ORG	PER
Training set	7140	3438	6321	6600
Development set	1837	922	1341	1842
Test set	1668	702	1661	1617

German data	LOC	MISC	ORG	PER
Training set	4363	2288	2427	2773
Development set	1181	1010	1241	1401
Test set	1035	670	773	1195

Table 2: Number of named entities per data file

Reuters Newswire + European Corpus Initiative

Sang and De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, Proceedings of CoNLL-2003

Summary of Results

- 16 systems participated
- Machine Learning Techniques
  - Combinations of Maximum Entropy Models (5) + Hidden Markov Models (4) + Winnow/Perceptron (4)
  - Others used once were Support Vector Machines, Conditional Random Fields, Transformation-Based learning, AdaBoost, and memory-based learning
  - Combining techniques often worked well
- Features
  - Choice of features is at least as important as ML method
  - Top-scoring systems used many types
  - No one feature stands out as essential (other than words)

	lex	pos	aff	pre	ort	gaz	chu	pat	cas	tri	bag	quo	doc
Florian	+	+	+	+	+	+	+	-	+	-	-	-	-
Chieu	+	+	+	+	+	+	-	-	-	+	-	+	+
Klein	+	+	+	+	-	-	-	-	-	-	-	-	-
Zhang	+	+	+	+	+	+	+	-	-	+	-	-	-
Carreras (a)	+	+	+	+	+	+	+	+	-	+	+	-	-
Curran	+	+	+	+	+	+	-	+	+	-	-	-	-
Mayfield	+	+	+	+	+	-	+	+	-	-	-	+	-
Carreras (b)	+	+	+	+	+	-	-	+	-	-	-	-	-
McCallum	+	-	-	-	+	+	-	+	-	-	-	-	-
Bender	+	+	-	+	+	+	+	-	-	-	-	-	-
Munro	+	+	+	-	-	-	+	-	+	+	+	-	-
Wu	+	+	+	+	+	+	-	-	-	-	-	-	-
Whitelaw	-	-	+	+	-	-	-	-	+	-	-	-	-
Hendrickx	+	+	+	+	+	+	+	-	-	-	-	-	-
De Meulder	+	+	+	-	+	+	+	-	+	-	-	-	-
Hammerton	+	+	-	-	-	+	+	-	-	-	-	-	-

Table 3: Main features used by the the sixteen systems that participated in the CoNLL-2003 shared task sorted by performance on the English test data. Aff: affix information (n-grams); bag: bag of words; cas: global case information; chu: chunk tags; doc: global document information; gaz: gazetteers; lex: lexical features; ort: orthographic information; pat: orthographic patterns (like Aa0); pos: part-of-speech tags; pre: previously predicted NE tags; quo: flag signing that the word is between quotes; tri: trigger words.

Sang and De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, Proceedings of CoNLL-2003

Use of External Information

- Improvement from using Gazeteers vs. unlabeled data nearly equal
- Gazeteers less useful for German than English (higher quality)

	G	U	E	English	German
Zhang	+	-	-	19%	15%
Florian	+	-	+	27%	5%
Hammerton	+	-	-	22%	-
Carreras (a)	+	-	-	12%	8%
Chieu	+	-	-	17%	-
Hendrickx	+	+	-	7%	5%
De Meulder	+	+	-	8%	3%
Bender	+	+	-	3%	6%
Curran	+	-	-	1%	-
McCallum	+	+	-	?	?
Wu	+	-	-	?	?

Table 4: Error reduction for the two development data sets when using extra information like gazetteers (G), unannotated data (U) or externally developed named entity recognizers (E). The lines have been sorted by the sum of the reduction percentages for the two languages.

Precision, Recall, and F-Scores

English test	Precision	Recall	$F_{\beta=1}$
* Florian	88.99%	88.54%	88.76±0.7
* Chieu	88.12%	88.51%	88.31±0.7
Klein	85.93%	86.21%	86.07±0.8
Zhang	86.13%	84.88%	85.50±0.9
Carreras (b)	84.05%	85.96%	85.00±0.8
Curran	84.29%	85.50%	84.89±0.9
Mayfield	84.45%	84.90%	84.67±1.0
Carreras (a)	85.81%	82.84%	84.30±0.9
McCallum	84.52%	83.55%	84.04±0.9
Bender	84.68%	83.18%	83.92±1.0
Munro	80.87%	84.21%	82.50±1.0
Wu	82.02%	81.39%	81.70±0.9
Whitelaw	81.60%	78.05%	79.78±1.0
Hendrickx	76.33%	80.17%	78.20±1.0
De Meulder	75.84%	78.13%	76.97±1.2
Hammerton	69.09%	53.26%	60.15±1.3
Baseline	71.91%	50.90%	59.61±1.2

German test	Precision	Recall	$F_{\beta=1}$
* Florian	83.87%	63.71%	72.41±1.3
* Klein	80.38%	65.04%	71.90±1.2
* Zhang	82.00%	63.03%	71.27±1.5
Mayfield	75.97%	64.82%	69.96±1.4
Carreras (b)	75.47%	63.82%	69.15±1.3
Bender	74.82%	63.82%	68.88±1.3
Curran	75.61%	62.46%	68.41±1.4
McCallum	75.97%	61.72%	68.11±1.4
Munro	69.37%	66.21%	67.75±1.4
Carreras (a)	77.83%	58.02%	66.48±1.5
Wu	75.20%	59.35%	66.34±1.3
Chieu	76.83%	57.34%	65.67±1.4
Hendrickx	71.15%	56.55%	63.02±1.4
De Meulder	63.93%	51.86%	57.27±1.6
Whitelaw	71.05%	44.11%	54.43±1.4
Hammerton	63.49%	38.25%	47.74±1.5
Baseline	31.86%	28.89%	30.30±1.3

* Not significantly different

Sang and De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, Proceedings of CoNLL-2003

Combining Results

- What happens if we combine the results of all of the systems?
  - Used a majority-vote of 5 systems for each set
  - English:

    F = 90.30 (14% error reduction of best system)
  - German:

    F = 74.17 (6% error reduction of best system)

MUC Redux

- Task: fill slots of templates
- MUC-4 (1992)
  - All systems hand-engineered
  - One MUC-6 entry used learning; failed miserably

AYACUCHO, 19 JAN 89 – TODAY TWO PEOPLE WERE WOUNDED WHEN A BOMB EXPLODED IN SAN JUAN BAUTISTA MUNICIPALITY. OFFICIALS SAID THAT SHINING PATH MEMBERS WERE RESPONSIBLE FOR THE ATTACK ... ... POLICE SOURCES STATED THAT THE BOMB ATTACK INVOLVING THE SHINING PATH CAUSED SERIOUS DAMAGES ... ...

Figure 1: Snippet of a MUC-4 document

0	MESSAGE: ID	TST3-MUC4-0014
1	MESSAGE: TEMPLATE	1
2	INCIDENT: DATE	19-JAN-89
3	INCIDENT: LOCATION	PERU: SAN JUAN BAUTISTA (MUNICIPALITY)
4	INCIDENT: TYPE	BOMBING
5	INCIDENT: STAGE OF EXECUTION	ACCOMPLISHED
6	INCIDENT: INSTRUMENT ID	"BOMB"
7	INCIDENT: INSTRUMENT TYPE	BOMB:"BOMB"
8	PERP: INCIDENT CATEGORY	TERRORIST ACT
9	PERP: INDIVIDUAL ID	"SHINING PATH MEMBERS"
10	PERP: ORGANIZATION ID	"SHINING PATH"
11	PERP: ORGANIZATION CONFIDENCE	SUSPECTED OR ACCUSED BY AUTHORITIES:"SHINING PATH"
12	PHYS TGT: ID	-
13	PHYS TGT: TYPE	-
14	PHYS TGT: NUMBER	-
15	PHYS TGT: FOREIGN NATION	-
16	PHYS TGT: EFFECT OF INCIDENT	SOME DAMAGE:"-"
17	PHYS TGT: TOTAL NUMBER	-
18	HUM TGT: NAME	-
19	HUM TGT: DESCRIPTION	"PEOPLE"
20	HUM TGT: TYPE	CIVILIAN:"PEOPLE"
21	HUM TGT: NUMBER	2:"PEOPLE"
22	HUM TGT: FOREIGN NATION	-
23	HUM TGT: EFFECT OF INCIDENT	INJURY:"PEOPLE"
24	HUM TGT: TOTAL NUMBER	-

Figure 2: Example of a MUC-4 template

MUC Redux

- Fast forward 12 years ... now use ML!
- Chieu et. al. show a machine learning approach that can do as well as most of the hand-engineered MUC-4 systems
  - Uses state-of-the-art:
    - Sentence segmenter
    - POS tagger
    - NER
    - Statistical Parser
    - Co-reference resolution
  - Features look at syntactic context
    - Use subject-verb-object information
    - Use head-words of NPs
  - Train classifiers for each slot type

Chieu, Hai Leong, Ng, Hwee Tou, & Lee, Yoong Keok (2003). Closing the Gap: Learning-Based Information Extraction Rivaling Knowledge-Engineering Methods, In (ACL-03).

Slot	VAg	VPa	V-Prep	N-Prep
Human Target	DIE(12)	KILL(2)	IDENTIFY-AS(47)	MURDER-OF(3)
Perpetrator Individual	KIDNAP(5)	IMPLICATE(17)	ISSUE-FOR(73)	WARRANT-FOR(64)
Physical Target	MONSERRAT(420)	DESTROY(1)	THROW-AT(32)	ATTACK-ON(11)
Perpetrator Organization	KIDNAP(16)	BLAME(25)	SUSPEND-WITH(87)	GUERRILLA-OF(31)
Instrument ID	EXPLODE(4)	PLACE(5)	EQUIP-WITH(31)	EXPLOSION-OF(17)

Table 1: The top-ranking feature for each group of features and the classifier of a slot

	TST3				TST4		
	R	P	F		R	P	F
GE	58	54	56	GE	62	53	57
GE-CMU	48	55	51	GE-CMU	53	53	53
UMASS	45	56	50	SRI	44	51	47
Alice-ME	46	51	48	Alice-ME	46	46	46
SRI	43	54	48	NYU	46	46	46
Alice-SVM	45	46	45	UMASS	47	45	46
Alice-DT	38	53	44	Alice-SVM	47	40	43
NYU	40	46	43	Alice-DT	41	46	43
UMICH	40	39	39	BBN	40	43	41
Alice-NB	45	34	39	Alice-NB	52	33	40
BBN	29	43	35	UMICH	36	34	35

Table 4: Accuracy of all slots on the TST3 and TST4 test set

Best systems took 10.5 person-months of hand-coding!

IE Techniques: Summary

- Machine learning approaches are doing well, even without comprehensive word lists
  - Can develop a pretty good starting list with a bit of web page scraping
- Features mainly have to do with the preceding and following tags, as well as syntax and word “shape”
  - The latter is somewhat language dependent
- With enough training data, results are getting pretty decent on well-defined entities
- ML is the way of the future!

IE Tools

- Research tools
  - Gate
    - http://gate.ac.uk/
  - MinorThird
    - http://minorthird.sourceforge.net/
  - Alembic (only NE tagging)
    - http://www.mitre.org/tech/alembic-workbench/