CSE 454
Advanced Internet Systems

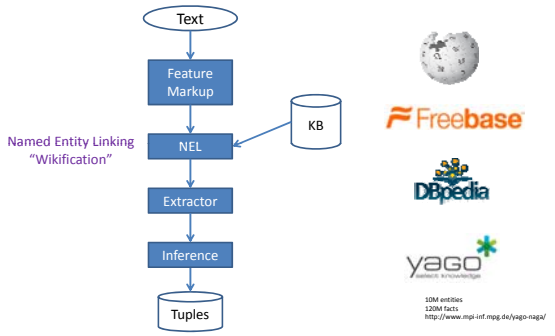
Slot-Filling Architectures

Dan Weld

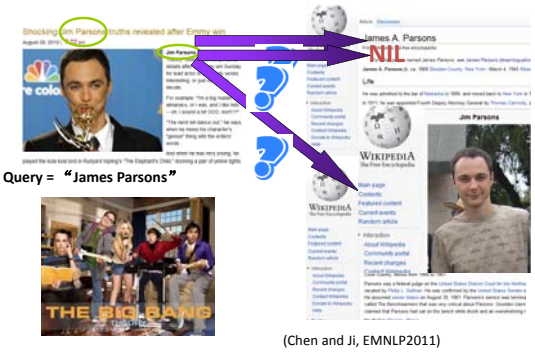
Logistics

- Datasets
  - Stored on Amazon
- Computing
  - $100 EC2 credit per student
  - You'll need to set up your own account
- Teams
  - 8-10 teams of 2-3 people
  - Bi-weekly meetings with Dan & Xiao
  - Additional mentor [optional]
- Classes
  - Steady state – one / week

High-Level Architecture



Entity Linking



Query = "James Parsons"

(Chen and Ji, EMNLP2011)

Slide from Dan Roth, Heng Ji, Taylor Cassidy, Quang Do TIE Tutorial

4

Slot Filling

Jim Parsons, a graduate of School Attended: University of Houston ...eater and Dance, won the Emmy on Sunday for Lead Actor in a Comedy Series for his work on The Big Bang Theory.

Slide from Dan Roth, Heng Ji, Taylor Cassidy, Quang Do TIE Tutorial

5

Named Entity Recognition *vs.* Linking

- NEL returns entity in KB which has been mentioned
- NER **identifies** *proper names* in texts, and **classifies** them into a set of predefined categories of interest.
  - Especially: **person**, **location** and **organisation**
  - Sometimes date/time expressions,
    measures (percent, money, weight etc),
    email addresses *etc.*
  - Sometimes domain-specific entities:
    names of drugs
    medical conditions,
    names of ships, *etc.*

Basic Problems in NE

- 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"

Example


Paris (disambiguation)

Places

Canada

United States

Other

People

Surname

Top Kurdish Militant Is Among Three Slain in Paris

PARIS — Three Kurdish women, including a founding member of a leading militant group fighting for autonomy in Turkey, were shot to death at a Kurdish institute in central Paris, police officials said on Thursday, potentially jeopardizing efforts to negotiate a cease-fire in the decades-old conflict.

News reports identified the women as Sakine Cansiz, a founder of the Kurdistan Workers' Party, known by the initials P.K.K.; Fidan Dogan, the head of the institute and a representative of the Kurdistan National Congress, an umbrella group of Kurdish organizations in Europe; and Leyla Soylemez, a young Kurdish

Battleships, movies, characters, paintings, songs…

Common NEL Features

Feature Category	Feature Type	Feature Description
Name	Spelling match	Exact string match, acronym match, alias match, string match based on edit distance, ratio of longest common subsequence to total string length, name component match, first letter match for abbreviations, organization suffix word match
	KB link mining	Name pairs mined from KB text redirect and disambiguation pages
	Name Gazetteer	Organization and geo-political entity abbreviation gazetteers

Common NEL Features

Feature Category	Feature Type	Feature Description
Name	Spelling match	Exact string match, acronym match, alias match, string match based on edit distance, ratio of longest common subsequence to total string length, name component match, first letter match for abbreviations, organization suffix word match
	KB link mining	Name pairs mined from KB text redirect and disambiguation pages
	Name Gazetteer	Organization and geo-political entity abbreviation gazetteers
Document surface	Lexical	Words in KB facts, KB text, query name, query text.
		Tf.idf of words and ngrams
	Position	Query name appears early in KB text
	Genre	Genre of the query text (newswire, blog, …)
	Local Context	Lexical and part-of-speech tags of context words
Entity Context	Type	Query entity type, subtype
	Relation	Entities co-occurred, or involved in some attributes/relations/events with the query
	Coreference	Coreference links between mentions in the source document and the KB text

Paris, France

Paris (English pronunciation: /ˈpærɪs/, /ˈpɛrɪs/; French pronunciation: [paʁi]) is the capital and largest city of France. It is situated on the river Seine in northern France, at the heart of the Île-de-France region. The city of Paris, within its administrative limits (the 20 arrondissements), has a population of about 2,230,000. Its metropolitan area is one of the largest population centres in Europe with more than 12 million inhabitants.

An important settlement for more than two millennia, Paris had become, by the 12th century, one of Europe's foremost centres of learning and the arts and the largest city in the Western world until the turn of the 18th century. Paris is today one of the world's leading business and cultural centres and its influences in politics, education, entertainment, media, science, and the arts all contribute to its status as one of the world's major global cities.

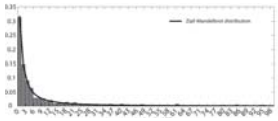
Paris and the Paris Region, with €572.4 billion in 2010, produce more than a quarter of the gross domestic product of France and has one of the largest city GDPs in the world.

Paris, Mythology

Paris (Ancient Greek: Πάρις; also known as **Alexander** or **Alexandros**, c.f. *Alaksandu* of *Wilusa*), the son of Priam, king of Troy, appears in a number of Greek legends. Probably the best-known was his elopement with Helen, queen of Sparta, this being one of the immediate causes of the Trojan War. Later in the war, he fatally wounds Achilles in the heel with an arrow, as foretold by Achilles's mother, Thetis.

Paris's childhood

Paris was a child of Priam and Hecuba (see the list of King Priam's children). Just before his birth, his mother dreamed that she gave birth to a flaming torch. This dream was interpreted by the seer Aesacus as a foretelling of the downfall of Troy, and he declared that the child would be the ruin of his homeland. On the day of Paris's birth it was further announced by Aesacus that the child born of a royal Trojan that day would have to be killed to spare the kingdom, being the child that would bring about the prophecy. Though Paris



Common NEL Features

Feature Category	Feature Type	Feature Description
Name	Spelling match	Exact string match, acronym match, alias match, string match based on edit distance, ratio of longest common subsequence to total string length, name component match, first letter match for abbreviations, organization suffix word match
	KB link mining	Name pairs mined from KB text redirect and disambiguation pages
	Name Gazetteer	Organization and geo-political entity abbreviation gazetteers
Document surface	Lexical	Words in KB facts, KB text, query name, query text.
		Tf.idf of words and ngrams
	Position	Query name appears early in KB text
	Genre	Genre of the query text (newswire, blog, …)
	Local Context	Lexical and part-of-speech tags of context words
Entity Context	Type	Query entity type, subtype
	Relation	Entities co-occurred, or involved in some attributes/relations/events with the query
	Coreference	Coreference links between mentions in the source document and the KB text
Profile		Slot fills of the query; KB attributes
Concept		Ontology extracted from KB text
Topic		Topics (identity and lexical similarity) for the query text and KB text
KB Link Mining		Attributes extracted from the hyperlink graphs (in-links, out-links) of the KB article
Popularity	Web	Top KB text ranked by search engine and its length
	Frequency	Frequency in KB texts

Joint Inference

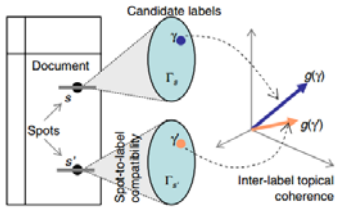
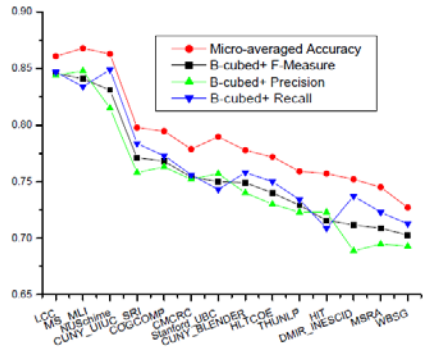


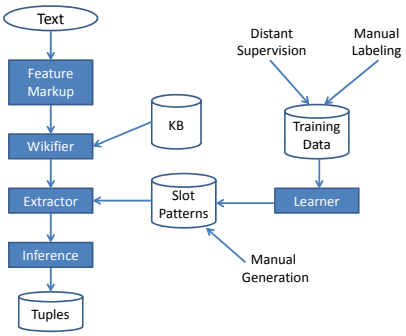
Figure 3: Labels $\gamma \in \Gamma_s, \gamma' \in \Gamma_{s'}$ have to be chosen for spots s, s' to maximize a combination of spot-to-label compatibility scores $\mathrm{NP}_s(\gamma), \mathrm{NP}_{s'}(\gamma')$ as well as topical similarity between γ and γ', say, $g(\gamma)^\top g(\gamma')$.

Kulkarni, S., Singh, A., Ramakrishnan, G., Chakrabarti, S., 2009. Collective annotation of Wikipedia entities in web text, in: Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining, pp. 457–466.

2011 Results NEL Monolingual



High-Level Architecture



Teams

- Named Entity Linking (1)
- Time (1)
- Distant Supervision (1)
- InstaRead (1)
- Relation-Specific (3-5)
- Lexicon Bootstrapping (0-1)

Information Extraction = Relation Extraction = Slot Filling

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…

NAME	TITLE	ORGANIZATION

Slides from Cohen & McCallum

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..

Slides from Cohen & McCallum

Landscape of IE Tasks (1/4): Pattern Feature Domain

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. 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

Non-grammatical snippets, rich formatting & links

Tables

Slides from Cohen & McCallum

Landscape of IE Tasks (2/4): Pattern Scope

Web site specific

Formatting

Amazon Book Pages

Genre specific

Layout

Resumes

Wide, non-specific

Language

University Names

Slides from Cohen & McCallum

Landscape of IE Tasks (3/4): Pattern Complexity

E.g. word patterns:

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.

Slides from Cohen & McCallum

Landscape of IE Tasks (4/4): Pattern Combinations

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

Slides from Cohen & McCallum

Landscape of IE Models

Lexicons

Abraham Lincoln was born in Kentucky.

member?

Alabama
Alaska
…
Wisconsin
Wyoming

Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.

Classifier

which class?

Sliding Window

Abraham Lincoln was born in Kentucky.

Classifier

which class?

Try alternate window sizes:

Boundary Models

Abraham Lincoln was born in Kentucky.

BEGIN

Classifier

which class?

BEGIN END BEGIN END

Finite State Machines

Abraham Lincoln was born in Kentucky.

Most likely state sequence?

Context Free Grammar

Abraham Lincoln was born in Kentucky.

NNP NNP V V P NP

NP PP

VP

VP

S

Most likely parse?

…and beyond

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

Slides from Cohen & McCallum

Supremacy of Machine Learning

- Machine learning is preferred approach to
  - Speech recognition, Natural language processing
  - Web search – result ranking
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology
  - Sensor networks
  - …
- This trend is accelerating
  - Improved machine learning algorithms
  - Improved data capture, networking, faster computers
  - Software too complex to write by hand
  - New sensors / IO devices
  - Demand for self-customization to user, environment

Space of ML Problems

Type of Supervision
(eg, Experience, Feedback)

What is Being Learned?

	Labeled Examples	Reward	Nothing
Discrete Function	Classification		Clustering
Continuous Function	Regression		
Policy	Apprenticeship Learning	Reinforcement Learning	

Classification

from data to discrete classes

Spam filtering

data

prediction

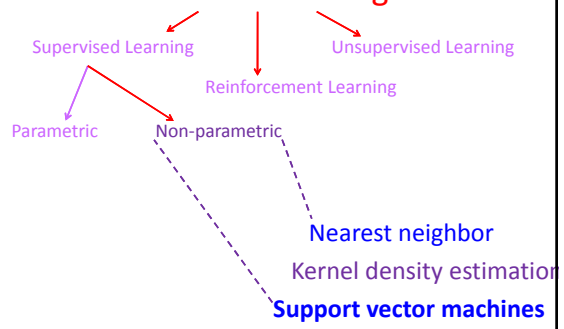
Weather prediction

The classification pipeline

Training

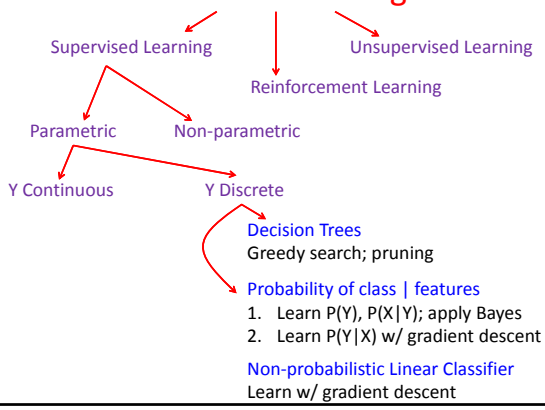
Testing

Machine Learning



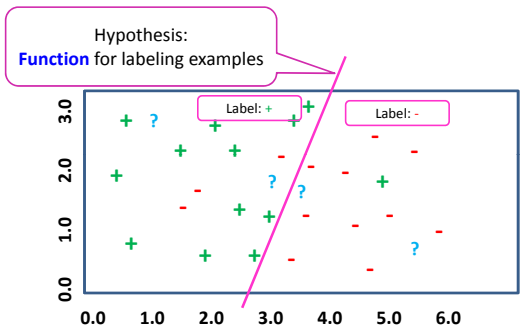
31

Machine Learning



32

Classifier



Terminology

- Examples
- Features
- Labels

Examples, Labels & Features for RE

\+ Citigroup has taken over EMI, the British …
\+ Citigroup's acquisition of EMI comes just ahead of …
\- Google's Adwords system has long included ways to connect to Youtube.

Terminology

- Examples
- Features
- Labels
- Training Sample
- Validation Sample
- Test Sample
- Loss Function
- Hypothesis Space

A Learning Problem

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Hypothesis Spaces

- **Complete Ignorance.** There are $2^{16} = 65536$ possible boolean functions over four input features. We can't figure out which one is correct until we've seen every possible input-output pair. After 7 examples, we still have 2^9 possibilities.

x_1	x_2	x_3	x_4	y
0	0	0	0	?
0	0	0	1	?
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	?
1	0	0	0	?
1	0	0	1	1
1	0	1	0	?
1	0	1	1	?
1	1	0	0	0
1	1	0	1	?
1	1	1	0	?
1	1	1	1	?

Knowledge-Based Weak Supervision

Heuristic match → training date → learn extractor

Acquisitions Database

Google	YouTube
Citigroup	EMI
Oracle	Sun

[Craven & Kumlien ICISMB-99]

Citigroup has taken over EMI, the British …

Citigroup's acquisition of EMI comes just ahead of …

Google's Adwords system has long included ways to connect to Youtube.

Matching WP Infoboxes WP Article Text

Clearfield County was created in 1804 from parts of Huntingdon and Lycoming Counties but was administered as part of Centre County until 1812.

Its county seat is Clearfield.

2,972 km² (1,147 mi²) of it is land and 17 km² (7 mi²) of it (0.56%) is water.

As of 2005, the population density was 28.2/km².

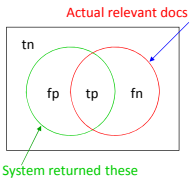
Precision & Recall

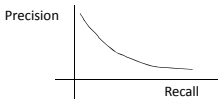
Precision $\frac{tp}{tp + fp}$

Percent of selected items that are correct

Recall $\frac{tp}{tp + fn}$

Percent of target items that were selected





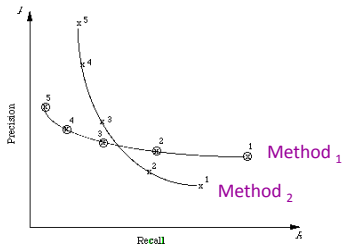
Precision-Recall Curve shows tradeoff

Precision/Recall Tradeoff

- Can get high precision (but low recall)
  - How?
- Can get high recall (but low precision)
  - How?
- Recall is a non-decreasing function of the number of docs retrieved
  - Precision usually decreases (in a good system)

Precision-Recall Curves

- May return any # of results ordered by similarity
- By varying numbers of docs (levels of recall)
  - Produce a ***precision-recall curve***



slide from Raghavan, Schütze, Larson

A combined measure: F

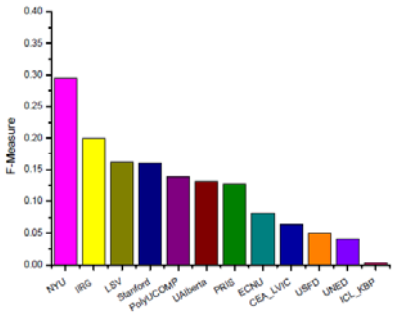
- Combined measure assessing tradeoff is ***F measure*** (weighted harmonic mean):

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

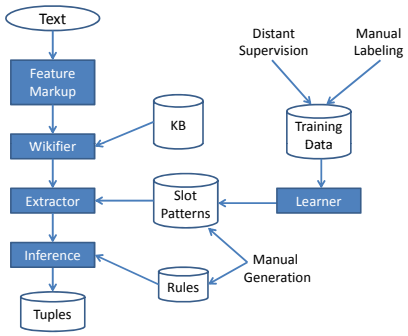
- People usually use balanced F_1 measure
  - i.e., with $\alpha = ½$
- Harmonic mean is conservative average
  - See CJ van Rijsbergen, *Information Retrieval*

slide from Raghavan, Schütze, Larson

2011 Slot-filling Results



High-Level Architecture

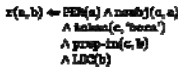


Rules are Crucial

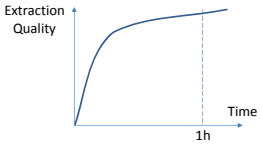
For both hand-engineered and learned systems

Goal: Enable experts to write quality rules *extremely quick*

- Example



- Rules as patterns
- Rules as features
- Rules (or space of rules) typically supplied

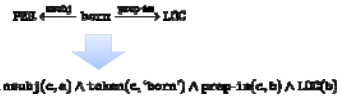


Desiderata and Challenges

1. User writes rules in simple, expressive rule language
2. User instantly sees rule extractions on large amounts of text

Writing Rules in Logic

- FOL[1] is simple, expressive, extensible, widely used
- Introduce predicates for NER, dependencies, ...

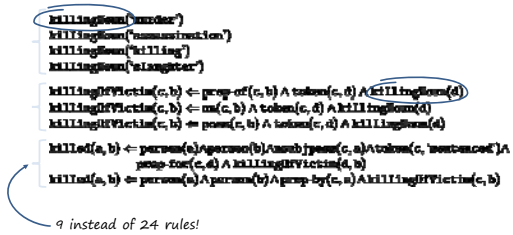


- Rules are deterministic, execute in defined order

[1] We use the subset referred to as 'safe domain-relational calculus'

Rule Composition

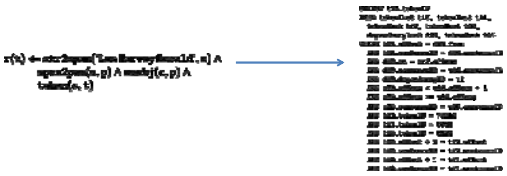
- Define new predicates for similar substructure



- More compact set of rules, better generalization

Enabling Instant Execution

- Translation to SQL allows dynamic optimization



- Each predicate maps to a fragment of SQL
- Intensional and extensional predicates
- Indices, caching, SSDs important

Desiderata and Challenges

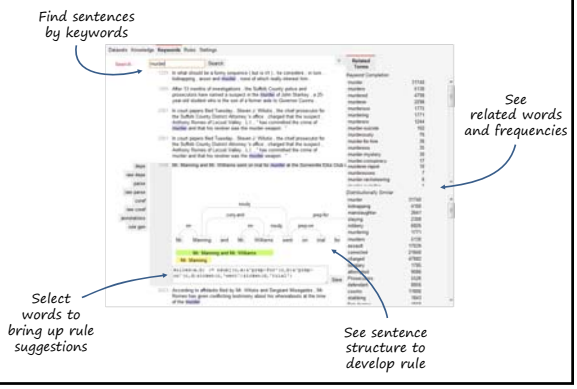
1. User writes rules in simple, expressive rule language
2. User instantly sees rule extractions on large amounts of text

Desiderata and Challenges

1. User writes rules in simple, expressive rule language
2. User instantly sees rule extractions on large amounts of text
3. User gets automatic rule suggestions based on distribution of data

Bootstrap Feature

Keyword Feature



Experimental Setup

Datasets		Procedure
22M news sentences	½ NYTimes[1]	**Develop** • 4 relational extractors in 55min each
1M news sentences	NYTimes07[1]	**Compare** • To a weakly supervised system
22M news sentences	½ NYTimes[1]	• Bootstrap, Keyword, Morphology, and Decomposition features
5K selected sentences (some gold annotations)	CoNLL04[2]	• To gold annotations for error analysis

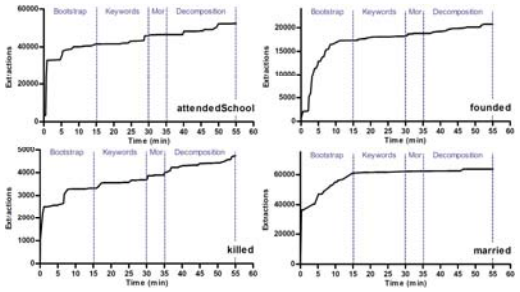
[1] LDC2008T19, [2] Roth & Yih, 2004

Comparison to Weakly Supervised Extraction

		attendedSchool	founded	killed	married
Rules	Precision	1.00	.91	.90	.90
	#extractions	1,411	997	189	4,694
Weakly supervised	Precision	0	.71	N/A	.50
	#extractions	5	14	N/A	2

- Precision consistently at least 90%
- Works well, even when weak supervision fails

Comparison of Development Phases



- Bootstrap initially effective, but recall limited
- Decomposition effective for some relations

Error Analysis

- On CoNLL04 'killed' wrt gold annotations: Re .34, Pr .98

False Negatives due to		False Positives due to	
Preprocessing (missing predictions)		Preprocessing (wrong predictions)	
NER	30	NER	0
Dependencies	25	Dependencies	2
Co-references	7	Co-references	0
Rules (missing predictions)		Rules (wrong predictions)	
Lexical items	89	Lexical items	0
Syntactic variation	17	Syntactic variation	0
Reasoning chain	10	Reasoning chain	0

- 36% of all errors are due to incorrect pre-processing