Information Extraction and Named Entity Recognition

- Slides from Christopher Manning
- http://web.stanford.edu/class/cs124/lec/Information_Extraction_and_Named_Entity_Recognition.ppt

Goal of Information Extraction

Extracting structured information out of unstructured text

Christopher Manning

Information Extraction

- Information extraction (IE) systems
  - Goal: produce a structured representation of relevant information:
    - *relations* (in the database sense), a.k.a.,
    - a *knowledge base*
  - Objectives:
    - Organize information so that it is useful to people
    - Put information in a semantically precise form that allows further inferences to be made by computer algorithms

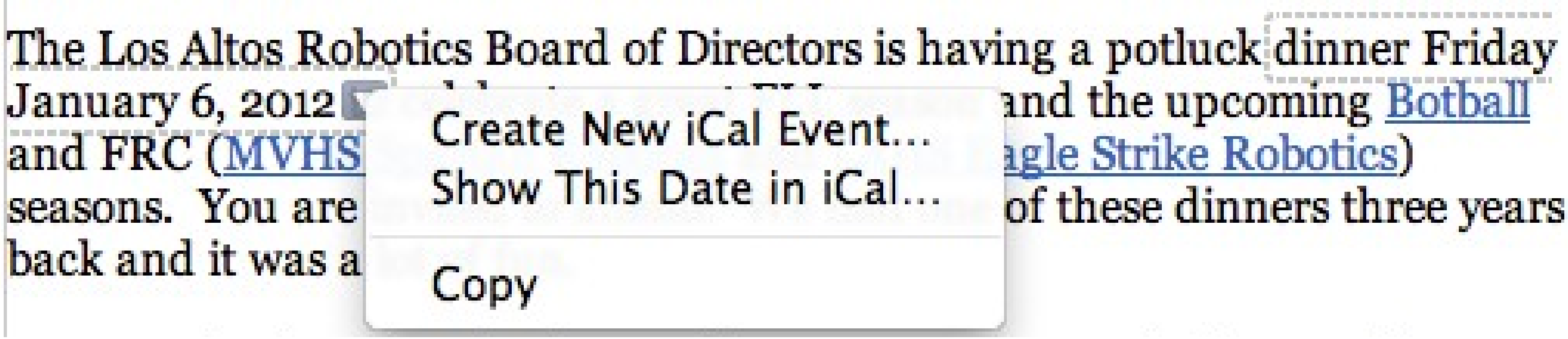
Information Extraction (IE)

- IE systems extract clear, factual information
  - Roughly: *Who did what to whom when?*
- E.g.,
  - Gathering earnings, profits, board members, headquarters, etc. from company reports
  - The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.
    - headquarters(“BHP Biliton Limited”, “Melbourne, Australia”)
  - Learn drug-gene product interactions from medical research literature
  - Extract citations from a research article

Christopher Manning

Low-level information extraction

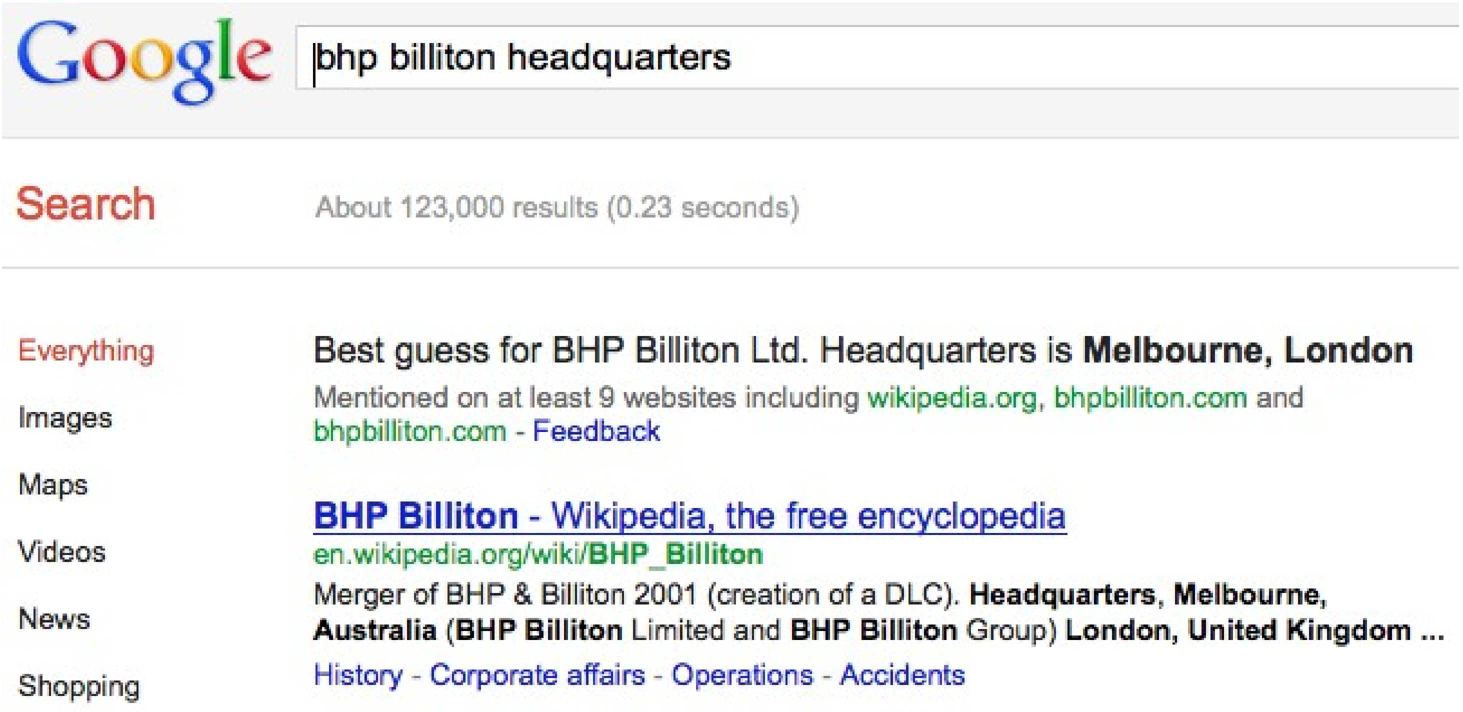
- Is now available – and I think popular – in applications like Apple or Google mail, and web indexing



- Often seems to be based on regular expressions and name lists

Christopher Manning

Low-level information extraction



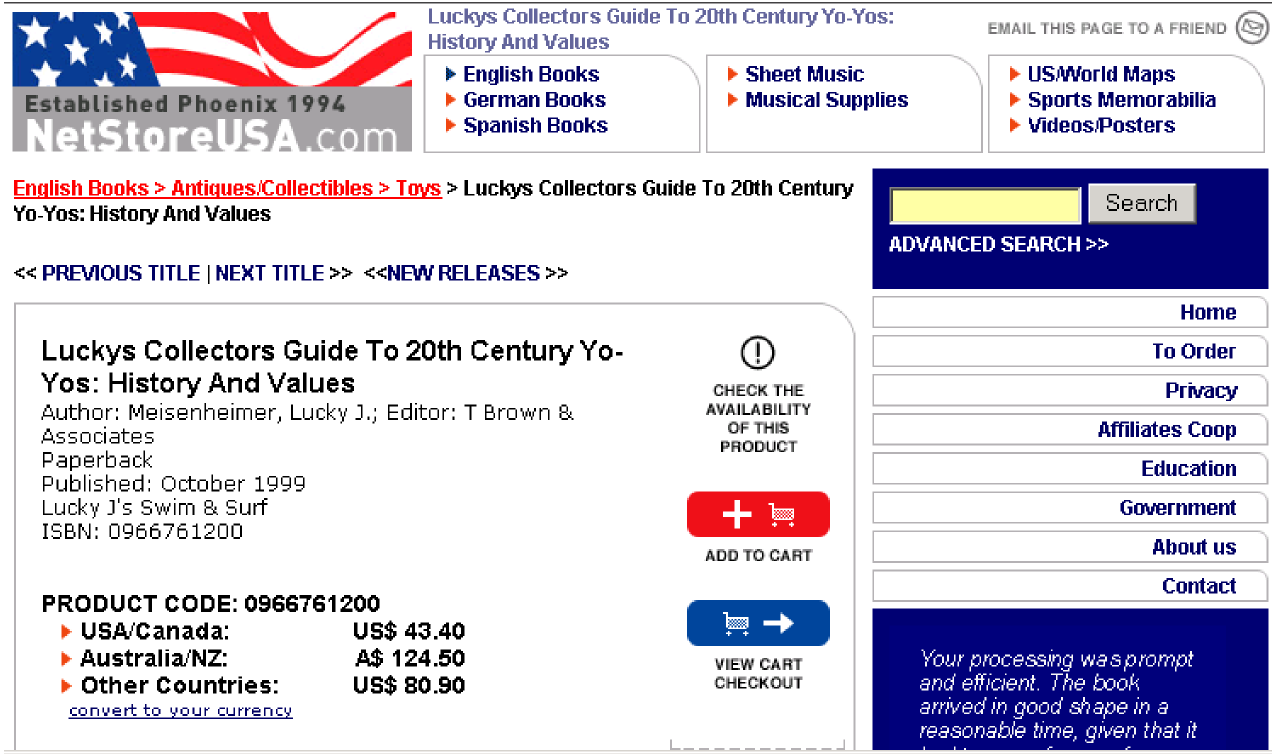
Christopher Manning

Why is IE hard on the web?

A book,
Not a toy

Title

Need this price



How is IE useful? Classified Advertisements (Real Estate)

Background:

- Plain text advertisements
- Lowest common denominator: only thing that 70+ newspapers using many different publishing systems can all handle

<ADNUM>2067206v1</ADNUM>
<DATE>March 02, 1998</DATE>
<ADTITLE>MADDINGTON $89,000</ADTITLE>
<ADTEXT>
OPEN 1.00 - 1.45
U 11 / 10 BERTRAM ST
NEW TO MARKET Beautiful
3 brm freestanding
villa, close to shops & bus
Owner moved to Melbourne
ideally suit 1st home buyer,
investor & 55 and over.
Brian Hazelden 0418 958 996
R WHITE LEEMING 9332 3477
</ADTEXT>

Christopher Manning

Why doesn't text search (IR) work?

What you search for in real estate advertisements:

- Town/suburb. You might think easy, but:
  - Real estate agents: Coldwell Banker, Mosman
  - Phrases: Only 45 minutes from Parramatta
  - Multiple property ads have different suburbs in one ad
- Money: want a range not a textual match
  - Multiple amounts: was $155K, now $145K
  - Variations: offers in the high 700s [*but not* rents for $270]
- Bedrooms: similar issues: br, bdr, beds, B/R

Christopher Manning

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Christopher Manning

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organi-
zation

Christopher Manning

Named Entity Recognition (NER)

- The uses:
  - Named entities can be indexed, linked off, etc.
  - Sentiment can be attributed to companies or products
  - A lot of IE relations are associations between named entities
  - For question answering, answers are often named entities.
- Concretely:
  - Many web pages tag various entities, with links to bio or topic pages, etc.
  - Reuters' OpenCalais, Evri, AlchemyAPI, Yahoo's Term Extraction, …
  - Apple/Google/Microsoft/… smart recognizers for document content

Evaluation of Named Entity Recognition

Precision, Recall, and the F measure;

their extension to sequences

Christopher Manning

The 2-by-2 contingency table

	correct	not correct
selected	tp	fp
not selected	fn	tn

Christopher Manning

Precision and recall

- **Precision**: % of selected items that are correct
  Recall: % of correct items that are selected

	correct	not correct
selected	tp	fp
not selected	fn	tn

The Named Entity Recognition Task

Task: Predict entities in a text

Foreign	ORG
Ministry	ORG
spokesman	O
Shen	PER
Guofang	PER
told	O
Reuters	ORG
:	:

Standard evaluation is per entity, *not* per token

Precision/Recall/F1 for IE/NER

- Recall and precision are straightforward for tasks like IR and text categorization, where there is only one grain size (documents)
- The measure behaves a bit funnily for IE/NER when there are *boundary errors* (which are *common*):
  - First Bank of Chicago announced earnings …
- This counts as both a fp and a fn
- Selecting *nothing* would have been better
- Some other metrics (e.g., MUC scorer) give partial credit (according to complex rules)

Christopher Manning

Three standard approaches to NER (and IE)

1. Hand-written regular expressions
   - Perhaps stacked
1. Using classifiers
   - Generative: Naïve Bayes
   - Discriminative: Maxent models
1. Sequence models
   - HMMs
   - CMMs/MEMMs
   - CRFs

Christopher Manning

Hand-written Patterns for Information Extraction

- If extracting from automatically generated web pages, simple regex patterns usually work.
  - Amazon page
  - <div class="buying"><h1 class="parseasinTitle">(.*?)</h1>
- For certain restricted, common types of entities in unstructured text, simple regex patterns also usually work.
  - Finding (US) phone numbers
  - (?:\(?[0-9]{3}\)?[-.])?[0-9]{3}[-.]?[0-9]{4}

Natural Language Processing-based Information Extraction

- For unstructured human-written text, some NLP may help
  - Part-of-speech (POS) tagging
  - Mark each word as a noun, verb, preposition, etc.

  - Syntactic parsing
  - Identify phrases: NP, VP, PP

  - Semantic word categories (e.g. from WordNet)
  - KILL: kill, murder, assassinate, strangle, suffocate

Rule-based Extraction Examples

Determining which person holds what office in what organization

- [person] , [office] *of* [org]
- Vuk Draskovic, leader of the Serbian Renewal Movement

- [org] (*named*, *appointed*, etc.) [person] Prep [office]
- NATO appointed Wesley Clark as Commander in Chief

Determining where an organization is located

- [org] *in* [loc]
- NATO headquarters in Brussels

- [org] [loc] (*division*, *branch*, *headquarters*, etc.)
- KFOR Kosovo headquarters

Information extraction as text classification

Naïve use of text classification for IE

- Use conventional classification algorithms to classify substrings of document as "*to be extracted*" or not.
- In some simple but compelling domains, this naive technique is remarkably effective.
  - But do think about when it would and wouldn't work!

Christopher Manning

‘Change of Address’ email

From: Robert Kubinsky <robert@lousycorp.com>
Subject: Email update

Hi all - I’m moving jobs and wanted to stay in touch
with everyone so....
My new email address is : robert@cubemedia.com
Hope all is well :)
>>R

Christopher Manning

Change-of-Address detection
[Kushmerick et al., ATEM 2001]

1. Classification

From: Robert Kubinsky <robert@lousycorp.com>
Subject: Email update

Hi all - I'm moving jobs and wanted to stay in touch
with everyone so....
My new email address is : robert@cubemedia.com
Hope all is well :)
>R

→ "message" naïve Bayes model → not CoA

↓ Yes

everyone so.... My new email address is: robert@cubemedia.com Hope all is well :) >

From: Robert Kubinsky <robert@lousycorp.com> Subject: Email update Hi all - I'm

2. Extraction

↓

"address" naïve-Bayes model →

P[robert@lousycorp.com] = 0.28

P[robert@cubemedia.com] = 0.72

Change-of-Address detection results [Kushmerick et al., ATEM 2001]

- Corpus of 36 CoA emails and 5720 non-CoA emails
  - Results from 2-fold cross validations (train on half, test on other half)
  - Very skewed distribution intended to be realistic
  - Note very limited training data: only 18 training CoA messages per fold
  - 36 CoA messages have 86 email addresses; old, new, and miscellaneous

	P	R	F1
Message classification	98%	97%	98%
Address classification	98%	68%	80%

Sequence Models for Named Entity Recognition

Christopher Manning

The ML sequence model approach to NER

Training

1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

Testing

1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities

Christopher Manning

Encoding classes for sequence labeling

	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
‘s	O	O
new	O	O
painting	O	O

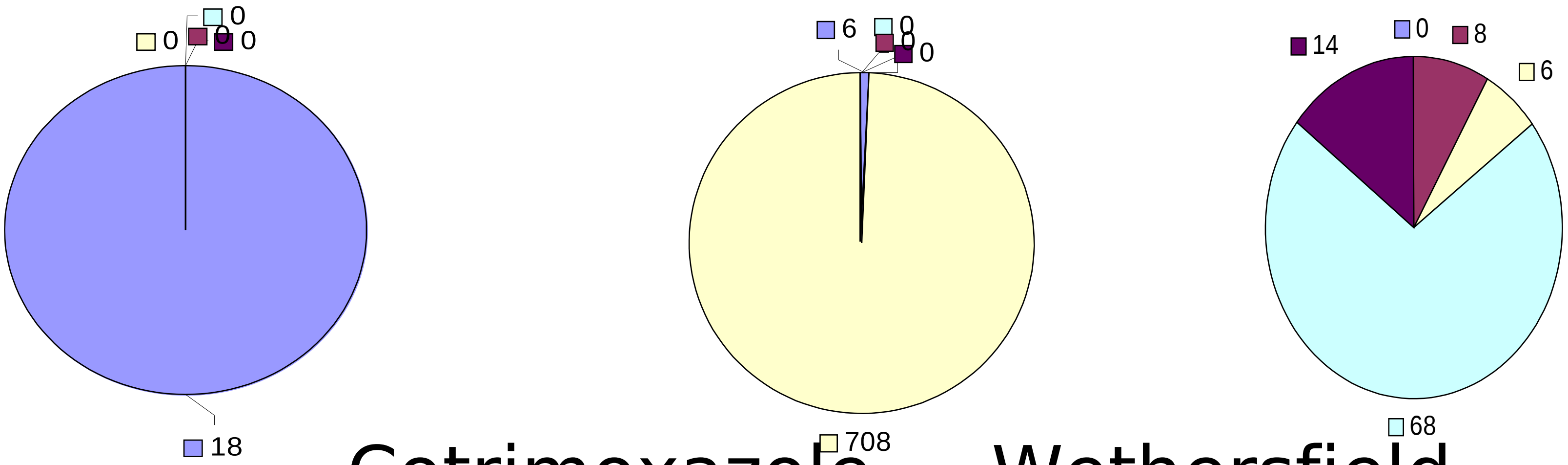
Features for sequence labeling

- Words
  - Current word (essentially like a learned dictionary)
  - Previous/next word (context)
- Other kinds of inferred linguistic classification
  - Part-of-speech tags
- Label context
  - Previous (and perhaps next) label

Christopher Manning

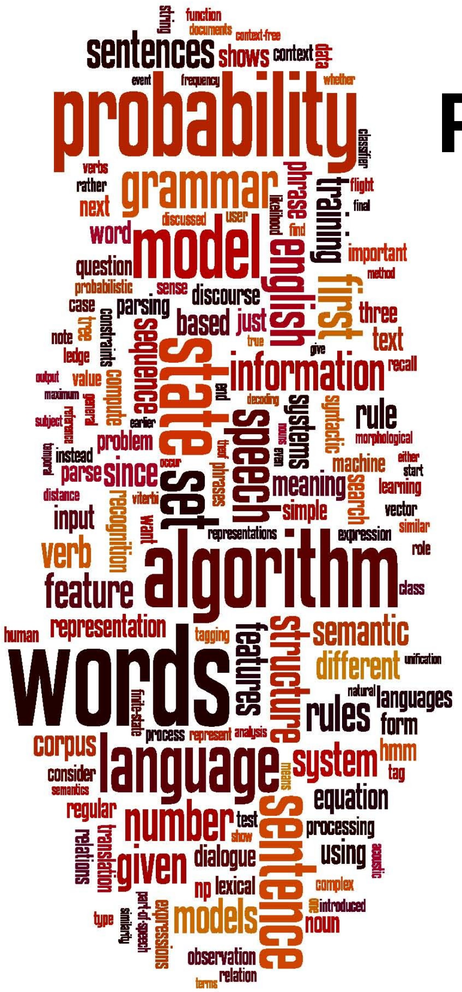
Features: Word substrings

oxa : field



Cotrimoxazole Wethersfield

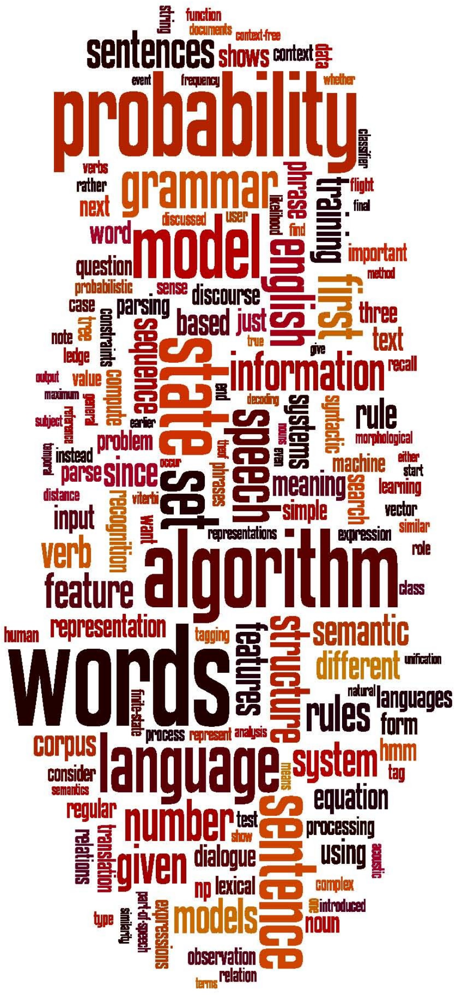
Alien Fury: Countdown to Invasion



Features: Word shapes

- Word Shapes
  - Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd



Relation Extraction

Christopher Manning

Binary Relation Association as Binary Classification

Christos Faloutsos conferred with Ted Senator, the KDD 2003 General Chair.

Person Person Role

Person-Role (Christos Faloutsos, KDD 2003 General Chair) → NO

Person-Role (Ted Senator, KDD 2003 General Chair) → YES

Christopher Manning

Resolving coreference (both within and across documents)

John Fitzgerald Kennedy was born at 83 Beals Street in Brookline, Massachusetts on Tuesday, May 29, 1917, at 3:00 pm,[7] the second s Joseph P. Kennedy, Sr., and Rose Fitzgerald; Rose, in turn, was the elde of John "Honey Fitz" Fitzgerald, a prominent Boston political figure who the city's mayor and a three-term member of Congress. Kennedy lived in Brookline for ten years and attended Edward Devotion School, Noble ar Greenough Lower School, and the Dexter School, through 4th grade. In the family moved to 5040 Independence Avenue in Riverdale, Bronx, N City; two years later, they moved to 294 Pondfield Road in Bronxville, N York, where Kennedy was a member of Scout Troop 2 (and was the firs Scout to become President).[8] Kennedy spent summers with his family at their home in Hyannisport, Massachusetts, and Christmas and Easter holidays with his family at their winter home in Palm Beach, Florida. For the 5th through 7th grade,

Christopher Manning

Rough Accuracy of Information Extraction

Information type	Accuracy
Entities	90-98%
Attributes	80%
Relations	60-70%
Events	50-60%

- IE tasks are hard!
- Errors cascade (error in entity tag → error in relation extraction)
- These are very rough, actually optimistic, numbers
  - Hold for well-established tasks, but lower for many specific/novel IE tasks