Chapter 8: Information Extraction (IE)

8.1 Motivation and Overview

8.2 Rule-based IE

8.3 Hidden Markov Models (HMMs) for IE

8.4 Linguistic IE

8.5 Entity Reconciliation

8.6 IE for Knowledge Acquisition

IE by text segmentation

Source: concatenation of structured elements with limited reordering and some missing fields

- Example: Addresses, bib records

House number	Building	Road	City	State	Zip
4089	Whispering Pines	Nobel Drive	San Diego	CA	92122

Author Year Title Journal Volume Page

P.P.Wangikar, T.P. Graycar, D.A. Estell, D.S. Clark, J.S. Dordick (1993) Protein and Solvent Engineering of Subtilising BPN' in Nearly Anhydrous Organic Media J.Amer. Chem. Soc. 115, 12231-12237.

Source: Sunita Sarawagi:
Information Extraction Using HMMs,
http://www.cs.cmu.edu/~wcohen/10-707/talks/sunita.ppt

8.3 Hidden Markov Models (HMMs) for IE

Idea:

text doc is assumed to be generated by a regular grammar (i.e. an FSA) with some probabilistic variation and uncertainty

→ **stochastic FSA = Markov model**

HMM – intuitive explanation:

• associate with each state a tag or symbol category (e.g. noun, verb, phone number, person name) that matches some words in the text;

• the instances of the category are given by a probability distribution of possible outputs in this state;

• the goal is to find a **state sequence** from an initial to a final state with **maximum probability of generating the given text**;

• the outputs are known, but the state sequence cannot be observed, hence the name ***hidden*** Markov model

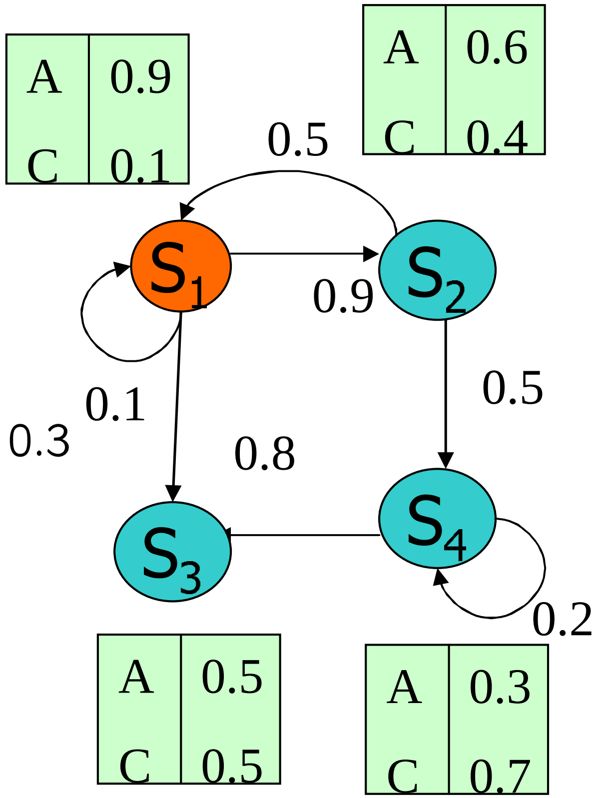
Hidden Markov Models in a Nutshell

- Doubly stochastic models

$\Pr(AACA) = \sum_{ijkl} \Pr(AACA, S_i S_j S_k S_l)$

$\Pr(AACA, S_i S_j S_k S_l) = \Pr(S_i)\Pr(A|S_i)\Pr(S_j|S_i)..\Pr(A|S_l)$

$\Pr(AACA, S_1 S_2 S_4 S_4) = 1 \times 0.9 \times 0.9 \times 0.6 \times 0.5 \times 0.7 \times 0.2 \times 0.3$



- Efficient dynamic programming algorithms exist for
  - Finding Pr(S)
  - The highest probability path P that maximizes Pr(S,P) (Viterbi)
- Training the model
  - (Baum-Welch algorithm)

Source: Sunita Sarawagi:
Information Extraction Using HMMs,
http://www.cs.cmu.edu/~wcohen/10-707/talks/sunita.ppt

Hidden Markov Model (HMM): Formal Definition

An HMM is a discrete-time, finite-state Markov model with
- state set $S = (s_1, ..., s_n)$ and the state in step t denoted $X(t)$,
- initial state probabilities p_i ($i=1, ..., n$),
- transition probabilities p_{ij}: $S\times S \rightarrow [0,1]$, denoted $p(s_i \rightarrow s_j)$,
- output alphabet $\Sigma = \{w_1, ..., w_m\}$, and
- state-specific **output probabilities q_{ik}: $S\times \Sigma \rightarrow [0,1]$**, denoted **$q(s_i \uparrow w_k)$** (or transition-specific output probabilities).

Probability of emitting output $o_1 ... o_k \in \Sigma^k$ is:

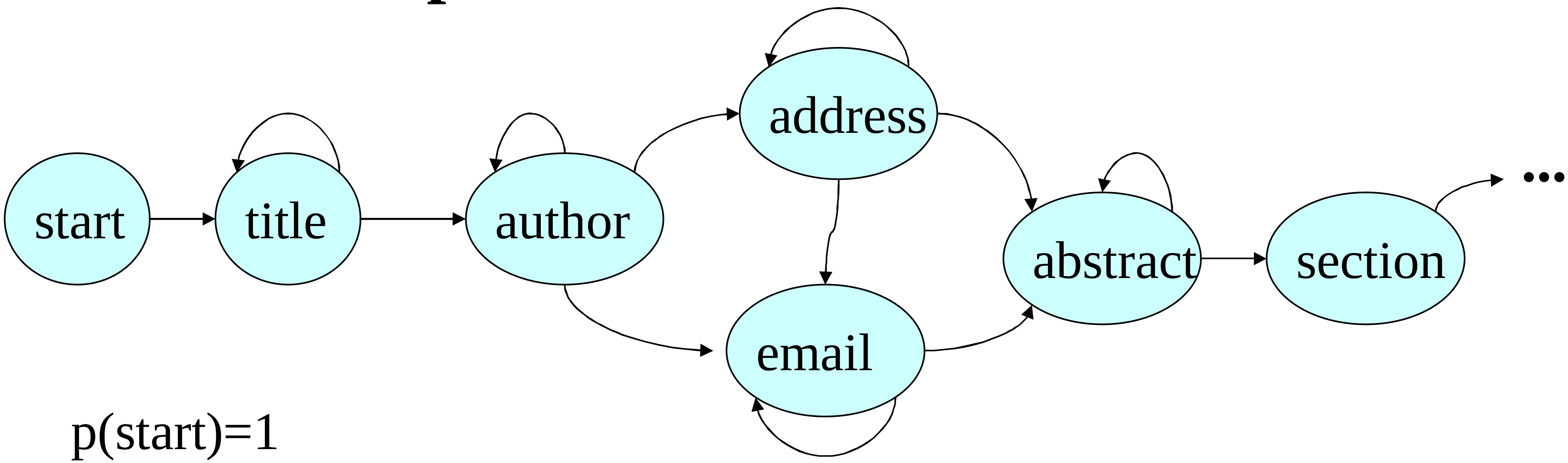
$$\sum_{x_1 ... x_k \in S} \prod_{i=1}^{k} p(x_{i-1} \rightarrow x_i)\, q(x_i \uparrow o_i) \quad \text{with} \quad p(x_0 \rightarrow x_1) := p(x_1)$$

can be computed iteratively with clever caching and reuse of intermediate results („memoization")

$$\alpha_i(t) := P[o_1 ... o_{t-1}, X(t) = i]$$

$$\alpha_i(1) = p(i) \qquad \alpha_j(t+1) = \sum_{i=1}^{n} \alpha_i(t)\, p(s_i \rightarrow s_j)\, p(s_i \uparrow o_t)$$

Example for Hidden Markov Model

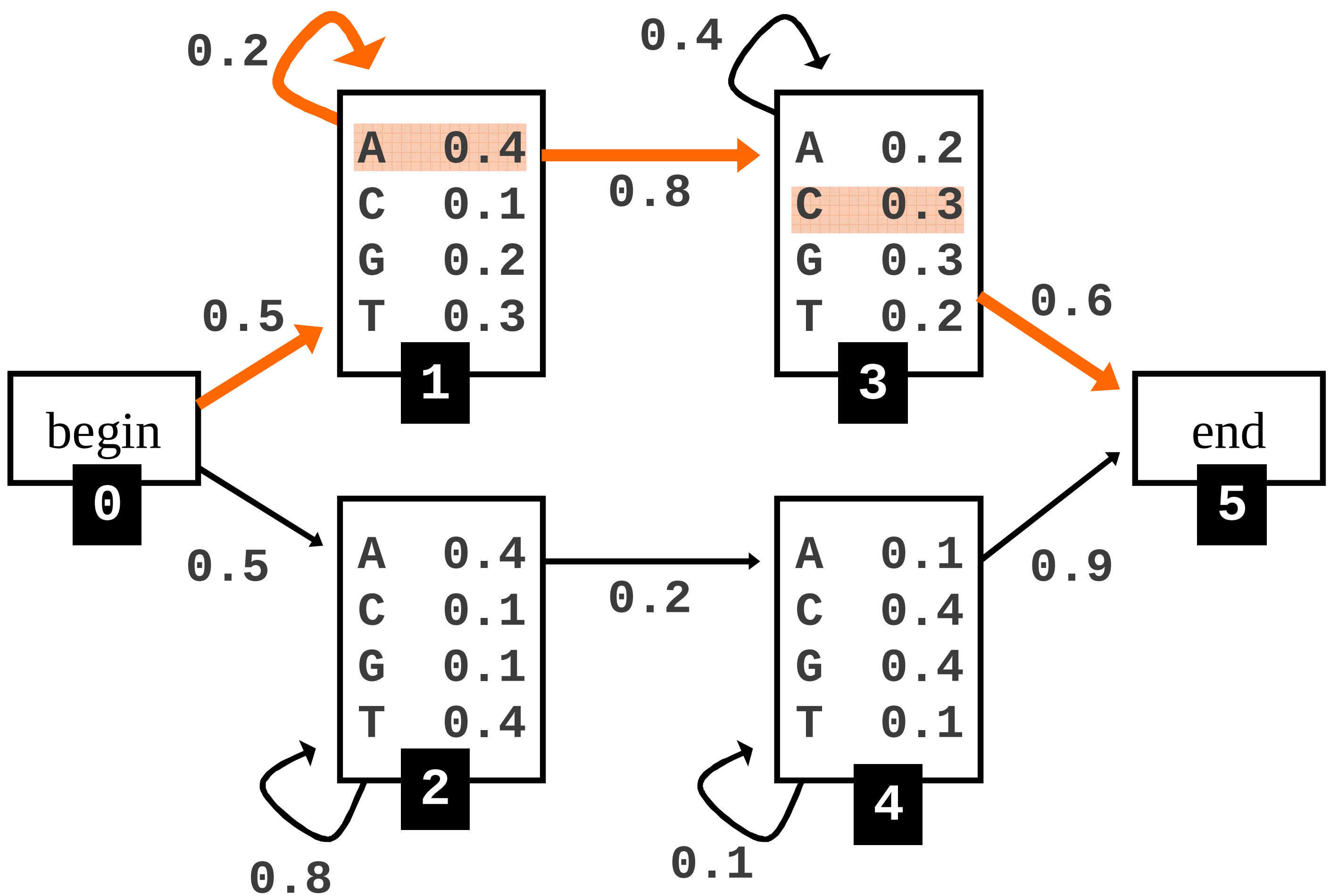


p(start)=1

p[author → author]=0.5
p[author → address]=0.2
p[author → email]=0.3
...

q[author ↑ <firstname>]= 0.1
q[author ↑ <initials>]= 0.2
q[author ↑ <lastname>]= 0.5
...
q[email ↑ @]=0.2
q[email ↑ .edu]=0.4
q[email ↑ <lastname>]=0.3
...

Example



$$\Pr(\text{AAC}, \pi) = a_{01} \times b_1(\text{A}) \times a_{11} \times b_1(\text{A}) \times a_{13} \times b_3(\text{C}) \times a_{35}$$

$$= 0.5 \times 0.4 \times 0.2 \times 0.4 \times 0.8 \times 0.3 \times 0.6$$

Source: Sunita Sarawagi: Information Extraction Using HMMs,
http://www.cs.cmu.edu/~wcohen/10-707/talks/sunita.ppt

Training of HMM

MLE for HMM parameters
(based on **fully tagged training sequences**)

$$p(s_i \rightarrow s_j) = \frac{\#\text{transitions } s_i \rightarrow s_j}{\sum_x \#\text{transitions } s_i \rightarrow x}$$

$$q(s_i \rightarrow w_k) = \frac{\#\text{outputs } s_i \uparrow w_k}{\sum_o \#\text{outputs } s_i \rightarrow o}$$

or use special case of **EM (Baum-Welch algorithm)**
to incorporate **unlabeled data**
(training: output sequence only, state sequence unknown)

learning of HMM structure (#states, connections): some work, but very difficult

Viterbi Algorithm for the Most Likely State Sequence

Find $\arg\max_{x_1 \ldots x_t} P[\text{state sequence } x_1 \ldots x_t \mid \text{output } o_1 \ldots o_t]$

Viterbi algorithm (uses dynamic programming):

$$\delta_i(t) := \max_{x_1 \ldots x_{t-1}} P[x_1 \ldots x_{t-1}, o_1 \ldots o_{t-1}, X(t) = i]$$

$$\delta_i(1) = p(i)$$

$$\delta_j(t+1) = \max_{i=1,\ldots,n} \delta_i(t)\, p(s_i \to s_j)\, q(s_i \uparrow o_t)$$

store argmax in each step

HMMs for IE

The following 6 slides are from:

Sunita Sarawagi:
Information Extraction Using HMMs,
http://www.cs.cmu.edu/~wcohen/10-707/talks/sunita.ppt

Combining HMMs with Dictionaries

- Augment dictionary
  - Example: list of Cities
- Exploit functional dependencies
  - Example
    - Santa Barbara -> USA
    - Piskinov -> Georgia

<u>Example:</u>

2001 University Avenue, Kendall Sq. Piskinov, Georgia

House number	Road Name	Area	City	State
2001	University Avenue,	Kendall Sq.,	Piskinov,	Georgia

House number	Road Name	Area	City	Country
2001	University Avenue,	Kendall Sq.,	Piskinov,	Georgia

Combining HMMs with Frequency Constraints

- Including constraints of the form: the same tag cannot appear in two disconnected segments
  - Eg: Title in a citation cannot appear twice
  - Street name cannot appear twice
- Not relevant for named-entity tagging kinds of problems

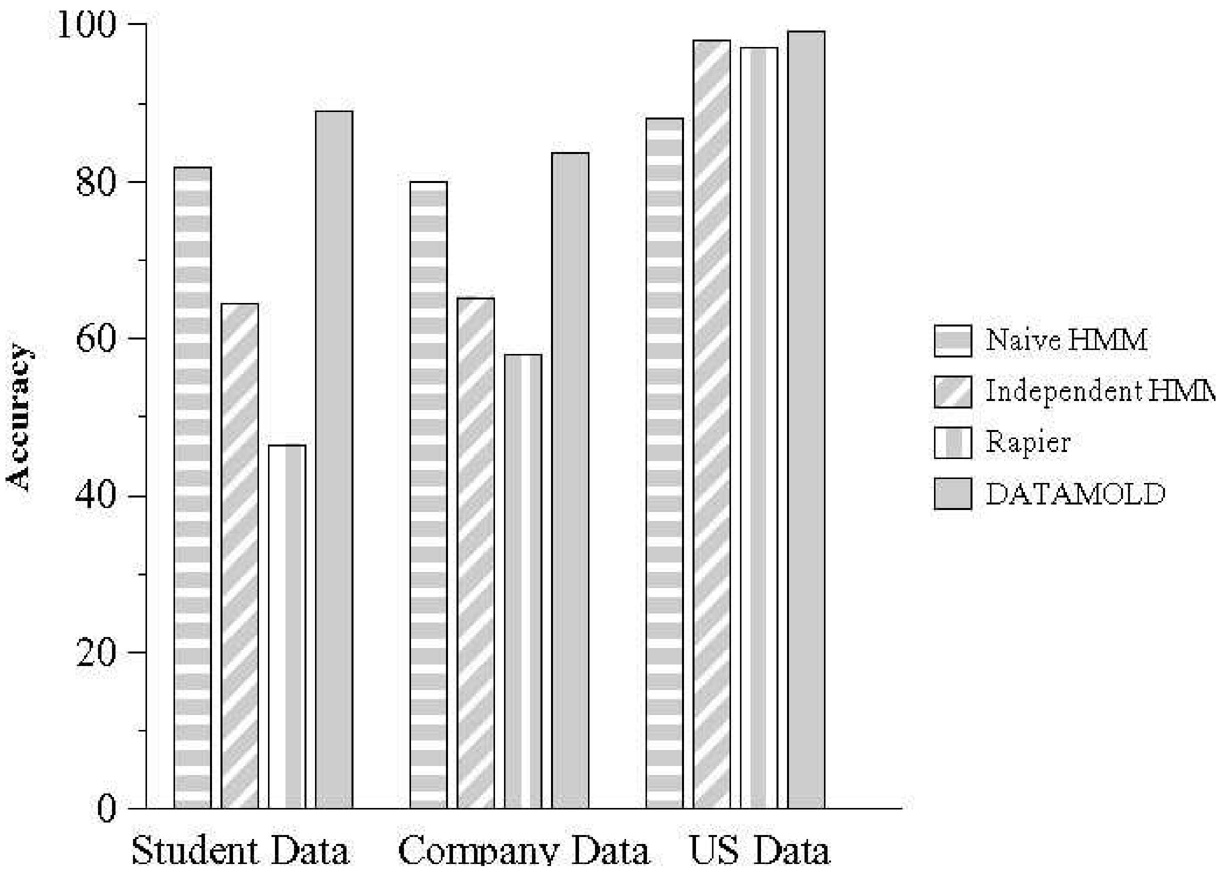
→ extend Viterbi algorithm with constraint handling

Comparative Evaluation

- Naïve model – One state per element in the HMM
- Independent HMM – One HMM per element;
- Rule Learning Method – Rapier
- Nested Model – Each state in the Naïve model replaced by a HMM

Results: Comparative Evaluation

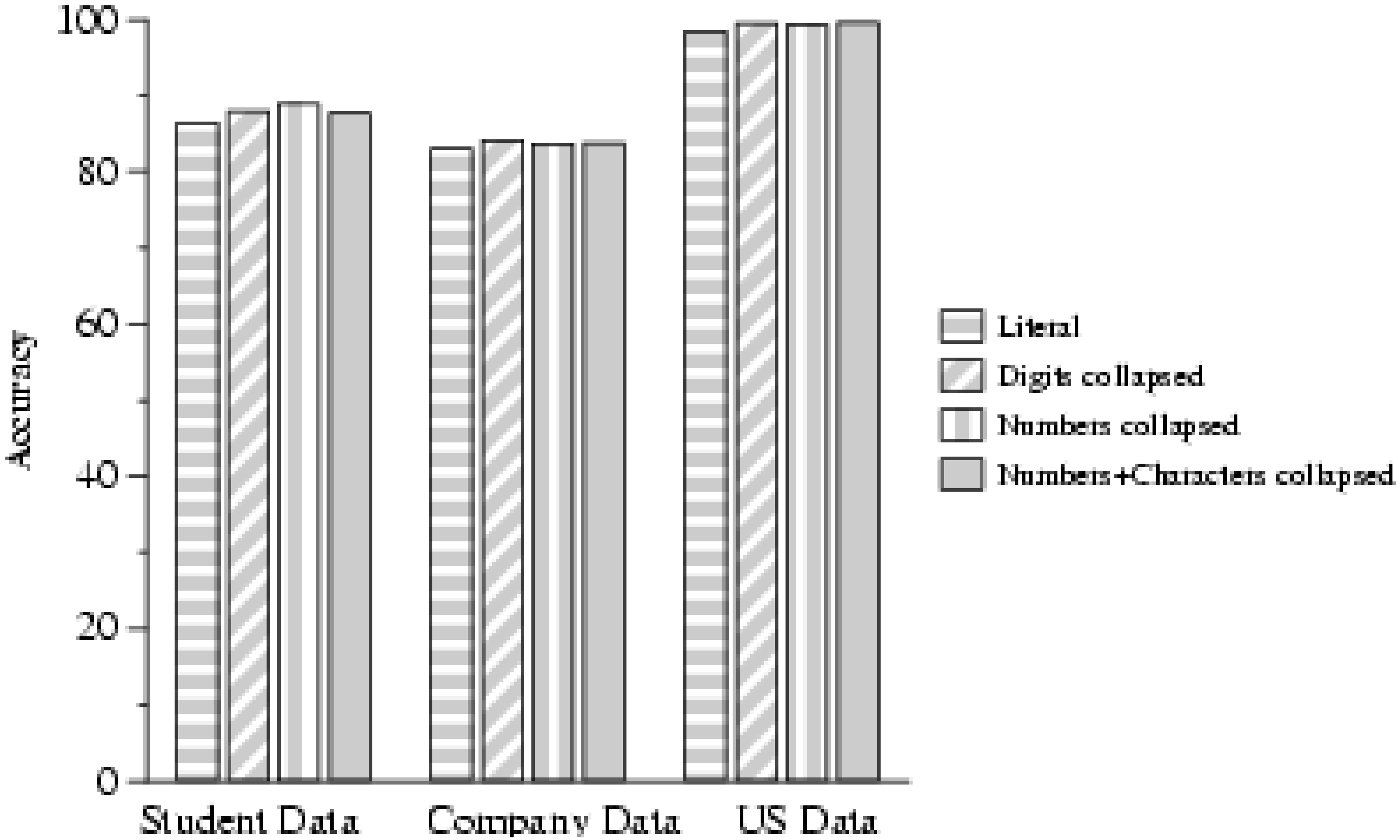
Dataset	instances	Elements
IITB student Addresses	2388	17
Company Addresses	769	6
US Addresses	740	6



The Nested model does best in all three cases

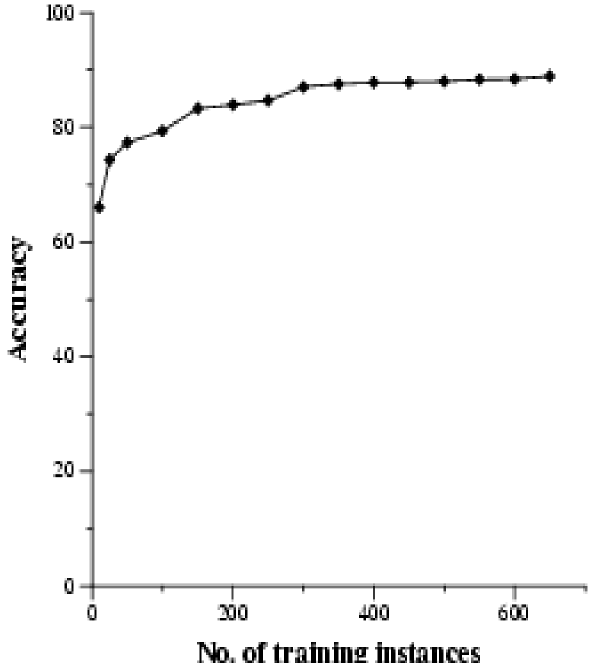
(from Borkar 2001)

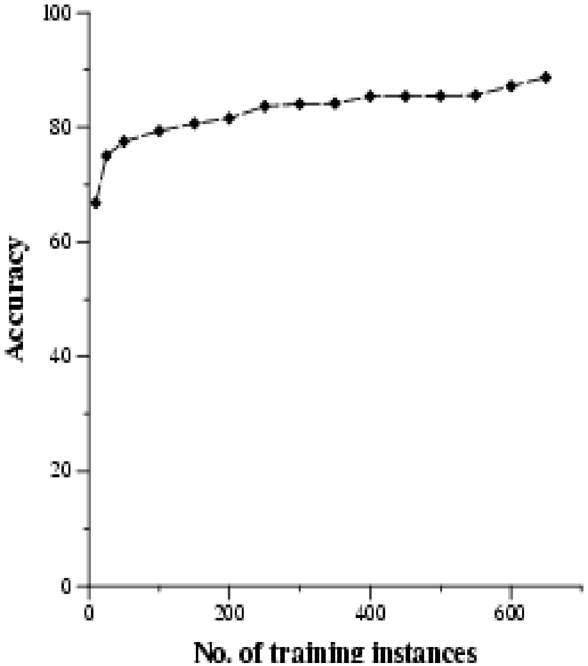
Results: Effect of Feature Hierarchy

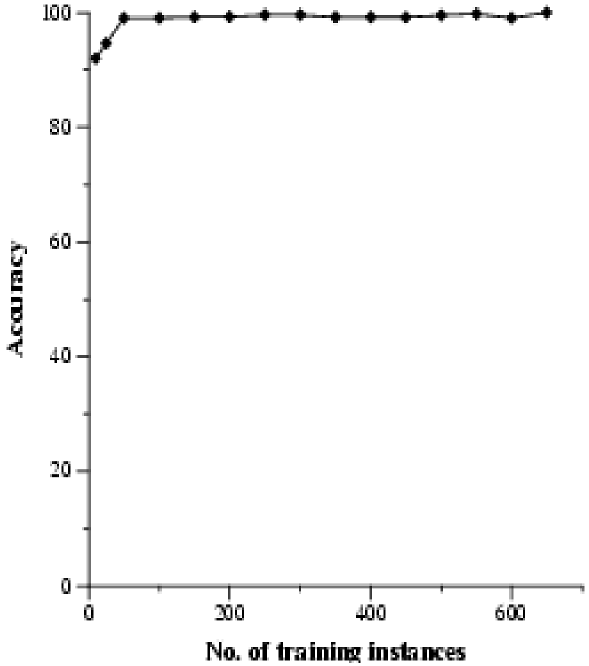


Feature Selection showed at least a 3% increase in accuracy

Results: Effect of training data size







HMMs are fast Learners.

We reach very close to the maximum accuracy with just 50 to 100 addresses

Semi-Markov Models for IE

The following 4 slides are from:

William W. Cohen
A Century of Progress on Information Integration:
a Mid-Term Report
http://www.cs.cmu.edu/~wcohen/webdb-talk.ppt

Features for information extraction

I met Prof. F. Douglas at the zoo

t	1	2	3	4	5	6	7	8
x	I	met	Prof	F.	Douglas	at	the	zoo.
y	**Other**	**Other**	**Person**	**Person**	**Person**	**other**	**other**	**Location**

$$\mathrm{f}(y_i, \mathrm{x}, i, y_{i-1})$$

Question: how can we guide this using a dictionary *D*?

Simple answer: make membership in *D* a feature f_d

Existing Markov models for IE

- Feature vector for each position

$$\mathbf{f}(y_i, \mathbf{x}, i, y_{i-1})$$

i-th label

Word i & neighbors

previous label

- Examples

$$f_2(y_i, \mathbf{x}, i, y_{i-1}) = 1 \text{ if } y_i \text{ is Person } \& \ x_i \text{ is Douglas}$$

$$f_3(y_i, \mathbf{x}, i, y_{i-1}) = 1 \text{ if } y_i \text{ is Person } \& \ y_{i-1} \text{ is Other}$$

- Parameters: weight *W* for each feature (vector)

Semi-markov models for IE

t	1	2	3	4	5	6	7	8
x	I	met	Prof.	F.	**Douglas**	at	the	zoo.
y	**Other**	**Other**	**Person**	**Person**	**Person**	**other**	**other**	**Location**

$$\mathbf{f}(y_i, \mathbf{x}, i, y_{i-1})$$

l,u	$l_1=u_1=1$	$l_2=u_2=2$	$l_3=3,\ u_3=5$			$l_4=6, u_4=6$	$l_5=u_5=7$	$l_6=u_6=8$
x	I	met	Prof.	F.	Douglas	at	the	zoo.
y	**Other**	**Other**	**Person**			**other**	**other**	**Location**

$$\mathbf{f}(y_j, \mathbf{x}, l_j, u_j, y_{j-1})$$

COST: Requires additional search in Viterbi
Learning and inference slower by O(maxNameLength)

Features for Semi-Markov models

$$\mathbf{f}(y_j, \mathbf{x}, l_j, u_j, y_{j-1})$$

j-th label

Start of S_j

end of S_j

previous label

$$f_2(y_j, \mathbf{x}, 3, 5, y_{j-1}) = 3 \quad \text{(segment length)}$$

$$f_3(y_j, \mathbf{x}, 3, 5, y_{j-1}) = 1 \text{ if } y_j \text{ is Person \& } y_{j-1} \text{ is Other}$$

$$f_5(P, \mathbf{x}, 3, 5, y_{j-1}) = 1 \text{ if}(x_3 x_4 x_5) = \text{Xx}_+\text{X.Xx}_+$$

$$f_4(P, \mathbf{x}, 3, 5, y_{j-1}) = \max_{e \in D} \text{cosine}(e, \text{"Prof.F.Douglas"})$$

Problems and Extensions of HMMs

- individual output letters/word may not show learnable patterns
  $\rightarrow$ output words can be entire **lexical classes**
  (e.g. numbers, zip codes)
- geared for flat sequences, not for structured text docs
  $\rightarrow$ use **nested HMM** where each state can hold another HMM
- cannot capture long-range dependencies
  (e.g. in addresses: with first word being „Mr.“ or „Mrs.“ the probability of later seeing a P.O. box rather than a street address decreases substantially)
  $\rightarrow$ use **dictionary lookups** in critial states and/or
  combine HMMs with other techniques for long-range effects
  $\rightarrow$ use **semi-Markov models**

8.4 Linguistic IE

<u>Preprocess input text using NLP methods:</u>

- Part-of-speech (PoS) tagging:
  each word (group) → grammatical role (NP, ADJ, VT, etc.)
- Chunk parsing: sentence → labeled segments (temp. adverb phrase, etc.)
- Link parsing: bridges between logically connected segments

<u>NLP-driven IE tasks:</u>

- Named Entity Recognition (NER)
- Coreference resolution (anaphor resolution)
- Template element construction
- Template relation construction
- Scenario template construction

…

- Logical representation of sentence semantics (e.g., FrameNet)

Named Entity Recognition and Coreference Resolution

Named Entity Recognition (NER):

- Run text through PoS tagging or stochastic-grammar parsing
- Use dictionaries to validate/falsify candidate entities

Example:

The shiny red rocket was fired on Tuesday. It is the brainchild of Dr. Big Head. Dr. Head is a staff scientist at We Build Rockets Inc.

→ <person> Dr. Big Head </person>
<person> Dr. Head </person>
<organization> We Build Rockets Inc </organization>
<time> Tuesday </time>

Coreference resolution (anaphor resolution):

- Connect pronous etc. to subject/object of previous sentence

Examples:

- The shiny red rocket was fired on Tuesday. It is the brainchild of Dr. Big Head.
  → … on Tuesday. It <reference> The shiny red rocket </reference> is the …
- Alas, poor Yorick, I knew him Horatio.

Template Construction

- Identify semantic relations of interest
  based on taxonomy of relations & classification
- Fill components of a tuple of an N-ary relation (slots of a frame)

<u>Example:</u>

Thompson is understood to be accused of importing heroin into the United States.

→ <event>

<type> drug-smuggling </type>

<destination> <country>United States</country></destination>

<source> unknown </unknown>

<perpetrator> <person> Thompson </person> </perpetrator>

<drug> heroin </drug>

</event>

Representation of extracted results:
FrameNet (625 different frame types)
or similar logic-based representation

very difficult; unclear if this works with decent accuracy

Logical Representation by FrameNet

Smuggling

Definition:

The words in this frame describe situations in which the Perpetrator secretly takes Goods into or out of a country or other area which are prohibited by law or on which one has not paid the required duty.

FEs:

Core:

Goal [Goal] — Goal is the location the Goods end up in.
Semantic Type
Goal

Goods [Goods] — The FE Goods is anything (including labor, time, or legal rights) that can be illegally take a country.

Path [Path] — The path refers to (a part of the) ground the Goods travel over or to a landmark the G

Perpetrator [Perp] — This is the person (or other agent) that illegally takes the goods into or out of a countr
Semantic Type
Sentient

Source [Src] — The source is the location the goods occupy initially before change of location.
Semantic Type
Source

Non-Core:

Duration [Dur] — The amount of time for which a state holds or a process is ongoing.
Semantic Type
Duration

Event [] — The unlawful movement of Goods.

Frequency [Freq] — The number of times that a smuggling event occurs.
Inmates frequently SMUGGLE marijuana into the prison

Manner [Man] — A description of the Event not covered by more specific FEs, including secondary effe *loudly*), and general descriptions comparing events (*the same way*). In most cases, it characteristics of a Perpetrator that also affect the action (*presumptuously, coldly, de eagerly, carefully*).
The rebels had secretly SMUGGLED in several tonnes of explosives.

Means [Mns] — An act of the Perpetrator which allows them to smuggle the Goods.

Place [Place] — Where the event takes place.
Semantic Type
Location

Purpose [Purp] — The action that the Perpetrator is trying to accomplish by the act of smuggling.
We SMUGGLED you in here to try to help but ...

Reason [Reas] — The Reason for which an event occurs.

Time [Time] — When the event occurs.
Semantic Type
Time

Inherits From: Committing_crime
Is Inherited By:
Subframe of:
Has Subframes:
Precedes:

Source:
http://framenet.icsi.berkeley.edu/

8.5 Entity Reconciliation (Fuzzy Matching, Entity Matching/Resolution, Record Linkage)

Problem:

- same entity appears in
  - different spellings (incl. mis-spellings, abbr., multilingual, etc.)
    e.g. Brittnee Speers vs. Britney Spears
    Microsoft Research vs. MS Research, Rome vs. Roma vs. Rom
  - different levels of completeness
    e.g. Britney Spears vs. Britney B. Spears
    Britney Spears (born Jan 1990) vs. Britney Spears (born 28/1/90)
    Microsoft (Redmond, USA) vs. Microsoft (Redmond, WA 98002)
- different entities happen to look the same
  e.g. George W. Bush vs. George W. Bush, Paris vs. Paris

- Problem even occurs within structured databases and requires data cleaning when integrating multiple databases (e.g. to build a data warehouse)
- Integrating heterogeneous databases or Deep-Web sources also requires schema matching

Entity Reconciliation Example

DB for Conference 1

PC

Name	Affiliation	Role
Alon Halevy Mike Franklin ...	U Washington UC Berkeley	... ...

Sessions

Title	Paper
XML Data	Unbreakable X Files

Papers

Paper	Title
P437	Unbreakable X Files

Authors

Paper	Name
Info Integration Dream Sensors Episode 1 ...	A. Halevy M.J. Franklin

DB for Conference 2

TrackChairs

Name	Organization
A. Halewi Michael J. Franklin ...	UW Seattle U California

Committee

Person	Org	Track
Sihem Amer-Yahia	AT&T	XML

PlenaryPapers

Paper	Session
Unbreakable Y	Beyond XML

AllPapers

Name	Authors	Session
Info Explosion Schema Bang ...	Halevy, ... M. Franklin	XML Era X Error

Entity Reconciliation: More Examples

The following 4 slides are from:

William W. Cohen:#

A Century of Progress on Information Integration:

A Mid-Term Report,

http://www.cs.cmu.edu/~wcohen/webdb-talk.ppt

Ted Kennedy's “Airport Adventure” [2004]

Washington -- Sen. Edward "Ted" Kennedy said Thursday that he was stopped and questioned at airports on the East Coast five times in March because his name appeared on **the government's secret "no-fly" list…**Kennedy was stopped because the name "**T. Kennedy**" has been used as an alias by someone on the list of terrorist suspects.

“…privately they [FAA officials] acknowledged being **embarrassed** that it took the senator and his staff more than three weeks to get his name removed.”

San Francisco Chronicle

Terror no-fly list singled out Kennedy
Senator was stopped 5 times at airports

Sara Kehaulani Goo, Washington Post

Friday, August 20, 2004

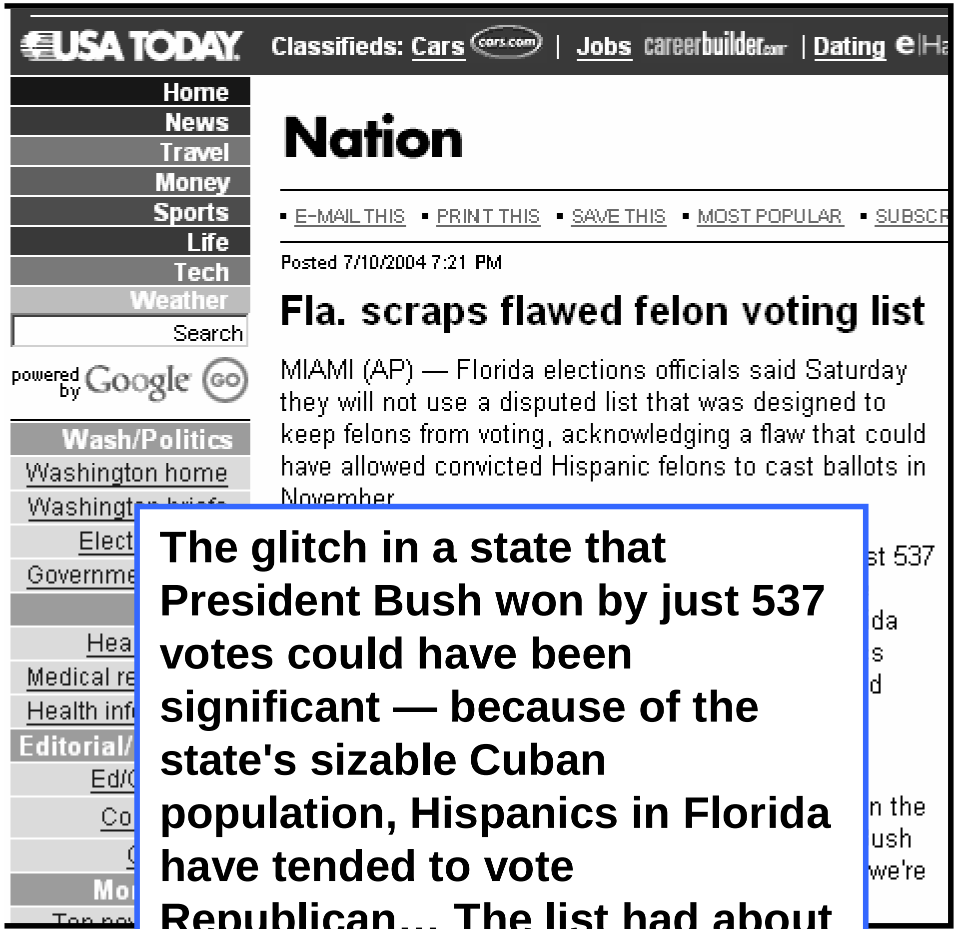
Washington -- Sen. Edward "Ted" Kennedy said Thursday that he was stopped and questioned at airports on the East Coast five times in March because his name appeared on the government's secret "no-fly" list.

- Printable Version
- Email This Article

Federal air security officials said the initial error that led to scrutiny of the Massachusetts Democrat should not have happened even though they recognize that the no-fly list is imperfect. But privately they acknowledged being embarrassed that it took the senator and his staff more than three weeks to get his name removed.

A senior administration official, who spoke on condition he not be identified, said Kennedy was stopped because the name "T. Kennedy" has been used as an alias by someone on the list of terrorist suspects.

Florida Felon List [2000, 2004]



USA TODAY Classifieds: Cars | Jobs | Dating

Home
News
Travel
Money
Sports
Life
Tech
Weather
Search

Wash/Politics
Washington home

Nation

E-MAIL THIS · PRINT THIS · SAVE THIS · MOST POPULAR

Posted 7/10/2004 7:21 PM

Fla. scraps flawed felon voting list

MIAMI (AP) — Florida elections officials said Saturday they will not use a disputed list that was designed to keep felons from voting, acknowledging a flaw that could have allowed convicted Hispanic felons to cast ballots in November.

The glitch in a state that President Bush won by just 537 votes could have been significant — because of the state's sizable Cuban population, Hispanics in Florida have tended to vote Republican… The list had about 28,000 Democrats and around 9,500 Republicans…

The purge of felons from voter rolls has been a thorny issue since the 2000 presidential election. A private company hired to identify ineligible voters before the election produced a list with scores of errors, and elections supervisors used it to remove voters without verifying its accuracy…

The new list … contained few people identified as Hispanic; of the nearly 48,000 people on the list created by the Florida Department of Law Enforcement, only 61 were classified as Hispanics.

Gov. Bush said the mistake occurred because two databases that were merged to form the disputed list were incompatible. … when voters register in Florida, they can identify themselves as Hispanic. But the potential felons database has no Hispanic category…

Matching University Courses

[Minton, Knoblock, et al 2001], [Doan, Domingos, Halevy 2001], [Richardson & Domingos 2003]

Goal might be to merge results of two IE systems:

Name:	Introduction to Computer Science
Number:	CS 101
Teacher:	M. A. Kludge
Time:	9-11am
Name:	Data Structures in Java
Room:	5032 Wean Hall

Title:	Intro. to Comp. Sci.
Num:	101
Dept:	Computer Science
Teacher:	Dr. Klüdge
TA:	John Smith
Topic:	Java Programming
Start time:	9:10 AM

When are two entities the same?

[1925]

- Bell Labs
- Bell Telephone Labs
- AT&T Bell Labs
- A&T Labs
- AT&T Labs—Research
- AT&T Labs Research, Shannon Laboratory
- Shannon Labs
- Bell Labs Innovations
- Lucent Technologies/Bell Labs Innovations

AT&T The world's networking company℠

History of Innovation: From 1925 to today, AT&T has attracted some of the world's greatest scientists, engineers and developers…. **[www.research.att.com]**

Lucent Technologies
Bell Labs Innovations

Bell Labs Facts: Bell Laboratories, the research and development arm of Lucent Technologies, has been operating continuously since 1925… [bell-labs.com]

Entity Reconciliation Techniques

- Edit distance measures (both strings and records)
- Exploit context information for higher-confidence matchings
  (e.g., publications and co-authors of Dave Dewitt vs. David J. DeWitt)
- Exploit reference dictionaries as ground truth
  (e.g. for address cleaning)
- Propagate matching confidence values
  in link-/reference-based graph structure
- Statistical learning in graph models

Additional Literature for Chapter 8

IE Overview Material:

- S. Chakrabarti, Section 9.1: Information Extraction
- N. Kushmerick, B. Thomas: Adaptive Information Extraction: Core Technologies for Information Agents, AgentLink 2003
- H. Cunningham: Information Extraction, Automatic, to appear in: Encyclopedia of Language and Linguistics, 2005, http://www.gate.ac.uk/ie/
- W.W. Cohen: Information Extraction and Integration: an Overview, Tutorial Slides, http://www.cs.cmu.edu/~wcohen/ie-survey.ppt
- S. Sarawagi: Automation in Information Extraction and Data Integration, Tutorial Slides, VLDB 2002, http://www.it.iitb.ac.in/~sunita/

Additional Literature for Chapter 8

Rule- and Pattern-based IE:

- M.E. Califf, R.J. Mooney: Relational Learning of Pattern-Match Rules for Information Extraction, AAAI Conf. 1999
- S. Soderland: Learning Information Extraction Rules fro Semi-Structured and Free Text, Machine Learning 34, 1999
- Arnaud Sahuguet, Fabien Azavant: Looking at the Web through XML Glasses, CoopIS Conf. 1999
- V. Crescenzi, G. Mecca: Automatic Information Extraction from
- Large Websites, JACM 51(5), 2004
- G. Gottlob, C. Koch, R. Baumgartner, M. Herzog, S. Flesca: The Lixto Data Extraction Project, PODS 2004
- A. Arasu, H. Garcia-Molina: Extracting Structured Data from Web Pages, SIGMOD 2003
- A. Finn, N. Kushmerick: Multi-level Boundary Classification for Information Extraction, ECML 2004

Additional Literature for Chapter 8

HMMs and HMM-based IE:

- Manning / Schütze, Chapter 9: Markov Models
- Duda/Hart/Stork, Section 3.10: Hidden Markov Models
- W.W. Cohen, S. Sarawagi: Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods, KDD 2004

Entity Rconciliation:

- W.W. Cohen: An Overview of Information Integration, Keynote Slides, WebDB 2005, http://www.cs.cmu.edu/~wcohen/webdb-talk.ppt
- S. Chaudhuri, R. Motwani, V. Ganti: Robust Identification of Fuzzy Duplicates, ICDE 2005

Knowledge Acquisition:

- O. Etzioni: Unsupervised Named-Entity Extraction from the Web: An Experimental Study, Artificial Intelligence 165(1), 2005
- E. Agichtein, L. Gravano: Snowball: extracting relations from large plain-text collections, ICDL Conf., 2000
- E. Agichtein, V. Ganti: Mining reference tables for automatic text segmentation, KDD 2004
- IEEE CS Data Engineering Bulletin 28(4), Dec. 2005, Special Issue on Searching and Mining Literature Digital Libraries