Automation in Information Extraction and Integration

Sunita Sarawagi

I I T Bombay

sunita@it.iitb.ac.in

Data integration

- The process of integrating data from multiple, heterogeneous, loosely structured information sources into a single well-defined structured database
- A tedious exercise involving
  - schema mapping,
  - structure/information extraction,
  - duplicate elimination,
  - missing value substitution,
  - error detection
  - standardization

Application scenarios

- Large enterprises:
  - Phenomenal amount of time and resources spent on data cleaning
  - Example: Segmenting and merging name-address lists during data warehousing
- Web:
  - Creating structured databases from distributed unstructured web-pages
    - Citation databases: Citeseer and Cora
- Other scientific applications
  - Bio-informatics
    - Extracting gene relations from medical text (KDD cup 2002)

Case study: CiteSeer

- Paper location:
  - Extract information from specific publisher websites
  - Extract ps/pdf files by searching the web with terms like “publications”
- Information extracted from papers:
  - Title, author from header
  - Extract citation entries
    - ➔ Bibliography section
    - ➔ Separate into individual records
    - ➔ Segment into title, author, date, page numbers etc
- Duplicate elimination across several citations to a paper (de-duplication)

Recent trends

- Classical problem that has bothered researchers and practitioners for decades
- Several existing commercial solutions for enterprise data integration [mid-80s]
  - Manual, domain-specific, data-driven script-based tools
  - Example: Name/address cleaning
  - Require high-expertise to code and maintain
- Desire to view "Web as a database" got machine learning researchers working on cleaning by learning from examples
- Several research prototypes, particularly in the context of web data integration

Scope of the tutorial

- Novel application of data mining and machine learning techniques to automate data cleaning operations.
- Distill recent research results from various areas:
  - Machine learning, data mining, information retrieval, natural language processing, web wrapper extraction
- Focus on two operations
  - Information Extraction
  - Duplicate elimination

Outline

- Information Extraction
  - Rule-based methods
  - Probabilistic methods
- Duplicate elimination
- Reducing the need for training data:
  - Active learning
  - Bootstrapping from structured databases
  - Semi-supervised learning
- Summary and research problems

Information Extraction (IE)

The IE task: Given,

- E: a set of structured elements (Target schema)
- S: unstructured source S

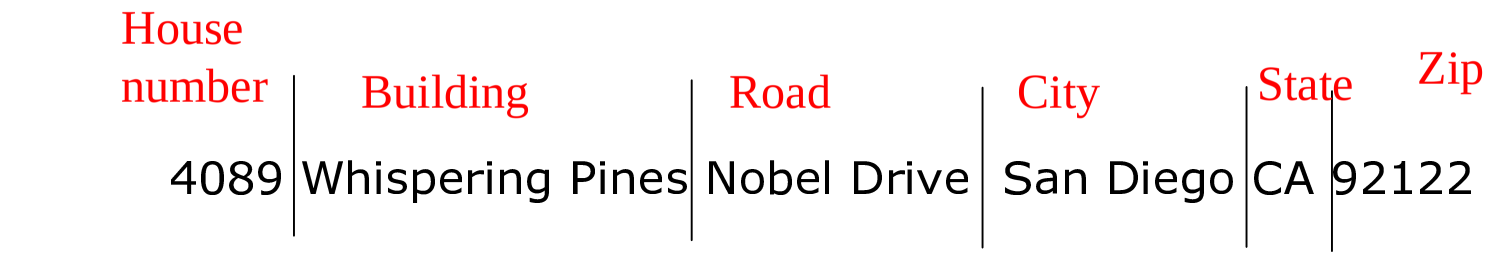
extract all instances of E from S

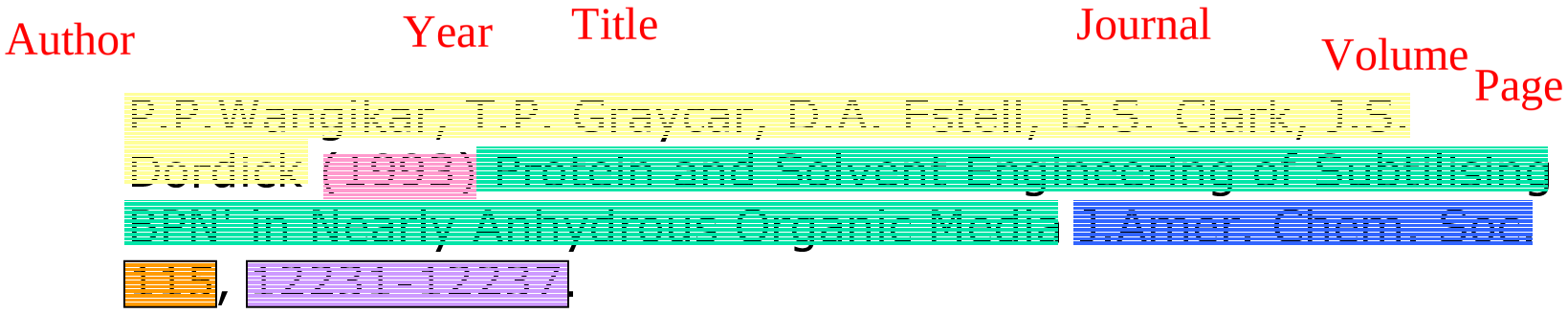
- Varying levels of difficulty depending on input and kind of extracted patterns
  - Text segmentation: Extraction by segmenting text
  - HTML wrapper: Extraction from formatted text
  - Classical IE: Extraction from free-format text

IE by text segmentation

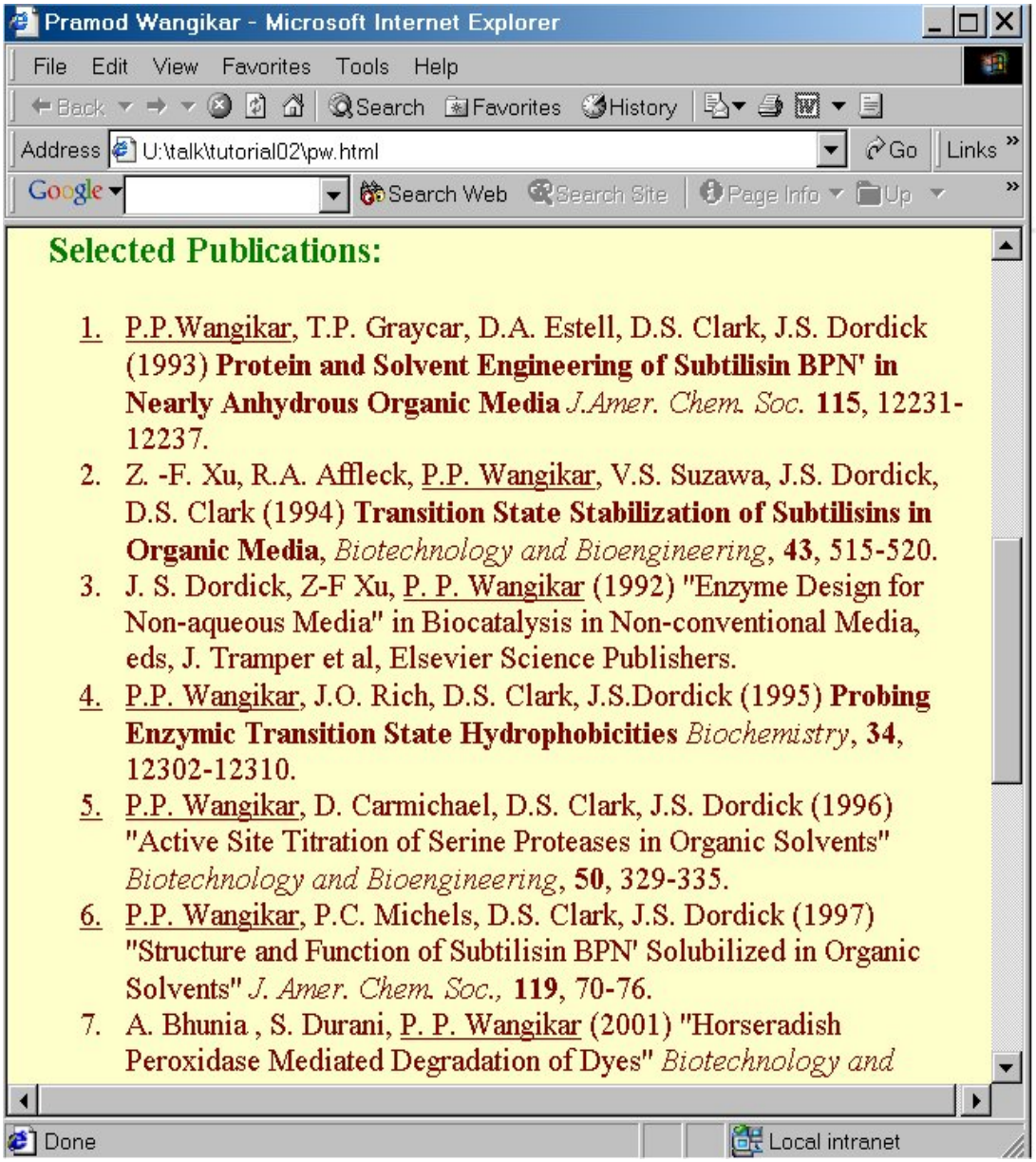
Source: concatenation of structured elements with limited reordering and some missing fields

- Example: Addresses, bib records





IE on formatted text: HTML wrappers



Source

<U><LI>P.P.Wangikar</U>, T.P. Graycar, D.A. Estell, D.S. Clark, J.S. Dordick (1993) <B>Protein and Solvent Engineering of Subtilisin BPN' in Nearly Anhydrous Organic Media</B> <I>J.Amer. Chem. Soc.</I> <B>115</B>, 12231-12237. </LI>

Output

AUTHOR: P.P.Wangikar..Dordick
YEAR: 1993
TITLE: Protein …. Media
JOURNAL: J.Amer. Chem. Soc
VOLUME: 115
PAGE: 12231-12237

HTML Wrappers

- Record level:
  - Extracting elements of a single list of homogeneous records from a page
  - Discovering record boundary by detecting regularity
- Page-level:
  - Extracting elements of multiple kinds of records
    - Example: name, courses, publications from home pages
- Site-level:
  - Example: populating a university database from pages of a university website

Research prototypes mostly at record and page level.

IE from free-format text

- Examples:
  - Gene interactions from medical articles
  - Part number, problem description from emails in help centers
  - Structured records describing an accident from insurance claims,
  - Merging companies, their roles and amount from news articles

Focus of NL researchers [Message Understanding Conferences (MUC)]

Requires deep linguistics and semantic analysis

We will discuss: Shallow IE based on syntactic cues

IE via machine learning

Given several examples showing position of structured elements in text,

Train a model to identify them in unseen text

At top-level a classification problem

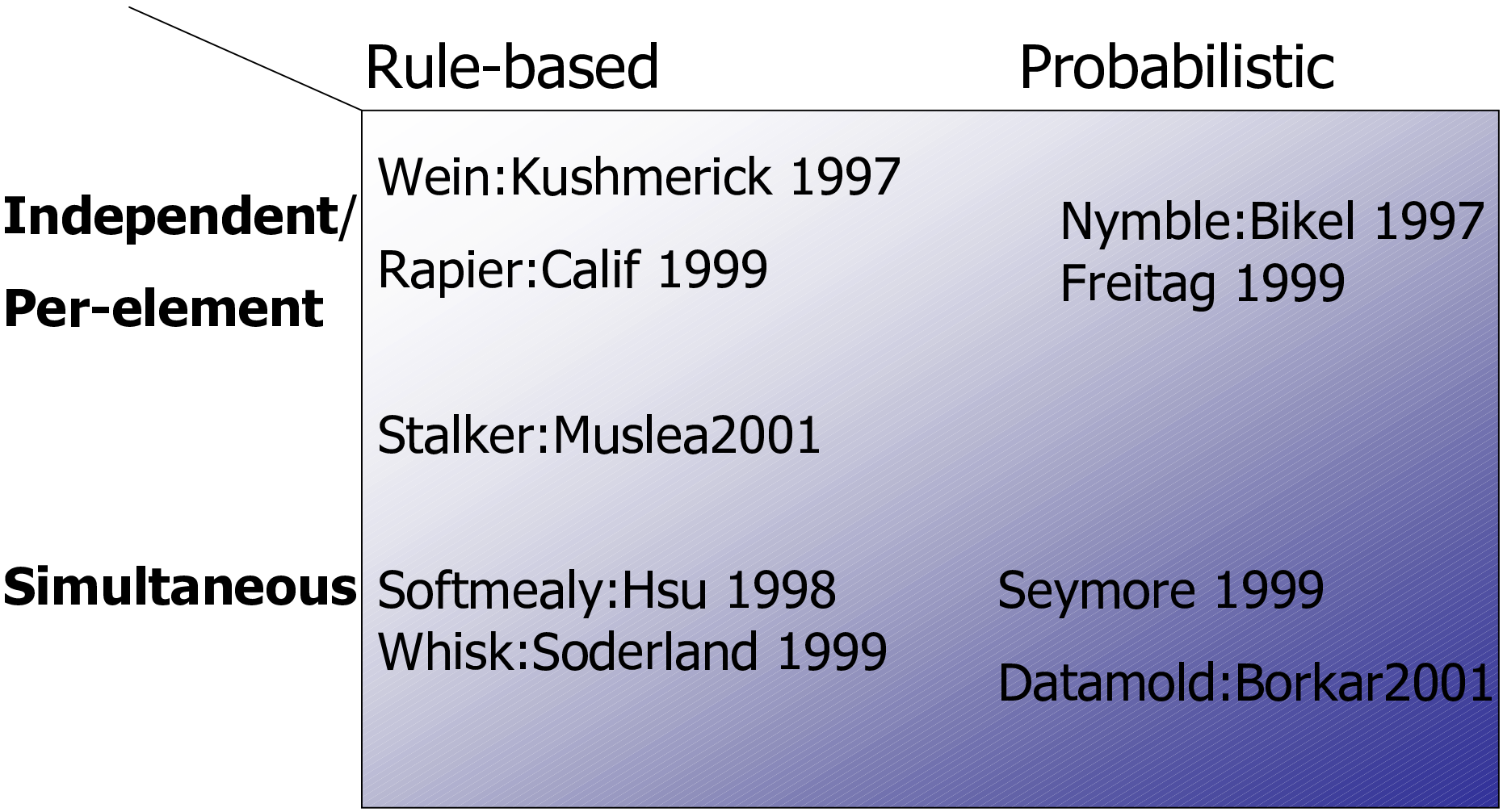
Issues:

- What are the input features?
- Build per-element classifiers or a single joint classifier?
- Which type of classifier to use?
- How much training data is required?
- Can one tell when the extractor is likely wrong?

Input features

- Content of the element
  - Specific keywords like street, zip, vol, pp,
  - Properties of words like capitalization, parts of speech, number?
- Formatting information: e.g., font, size
- Inter-element sequencing
- Intra-element sequencing
- Element length
- External database
  - Dictionary words
  - Semantic relationship between words
- Richer structure: tree, tables

Structure of IE models



Rule-based IE models

Stalker (Muslea et al 2001)

- Model type: Rules with conjuncts and disjuncts
  - For each element, two rules: start rules R1 and end rule R2
- Features:
  - html tags primarily
  - punctuations
  - predefined text features: isNumber, isCapitalized
- Relationship between elements:
  - Independent within same level of hierarchy
- Training method: basic sequential rule covering algorithm

Example

<LI> <U> P.P.Wangikar </U>, J.S. Dordick (1993) <B> Protein and Solvent Engineering of Subtilisin BPN' in Nearly Anhydrous Organic Media</B> <I>J.Amer. Chem. Soc.</I> <B>115</B>, 12231-12237. </LI>

<LI>A. Bhunia , S. Durani, <U>P. P. Wangikar</U> (2001) "Horseradish Peroxidase Mediated Degradation of Dyes" <I>Biotechnology and Bioengineering,</I> <B>72</B>, 562-567. </LI>

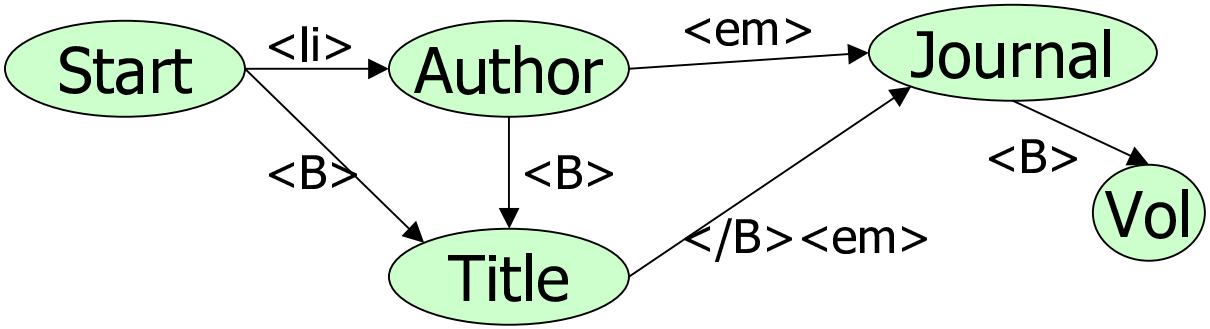
- Author:
  - R1: skipTo( )
  - R2: skipTo(( )
- Title:
  - R1: skipTo(<B>) OR skipTo(“) → disjunction
  - R2: skipTo(</B>) OR skipTo(“)
- Volume:
  - R1: skipTo(<B>) skipuntil(Number) → conjunction
  - R2: skipTo(</B>)

Limitations of rule-based approach

- As in WEIN, Stalker
  - No ordering dependency between elements
  - Non-overlap of elements not exploited
  - Position information ignored
  - Content largely ignored
  - Heuristics to order rule firing and ordering

Finite state machines

- Model ordering relationship between elements (Softmealy, Hsu 1998)
  - Node: elements to be extracted
  - Transition edge: rules marking start of element.
    - Rules are similar to those in STALKER.

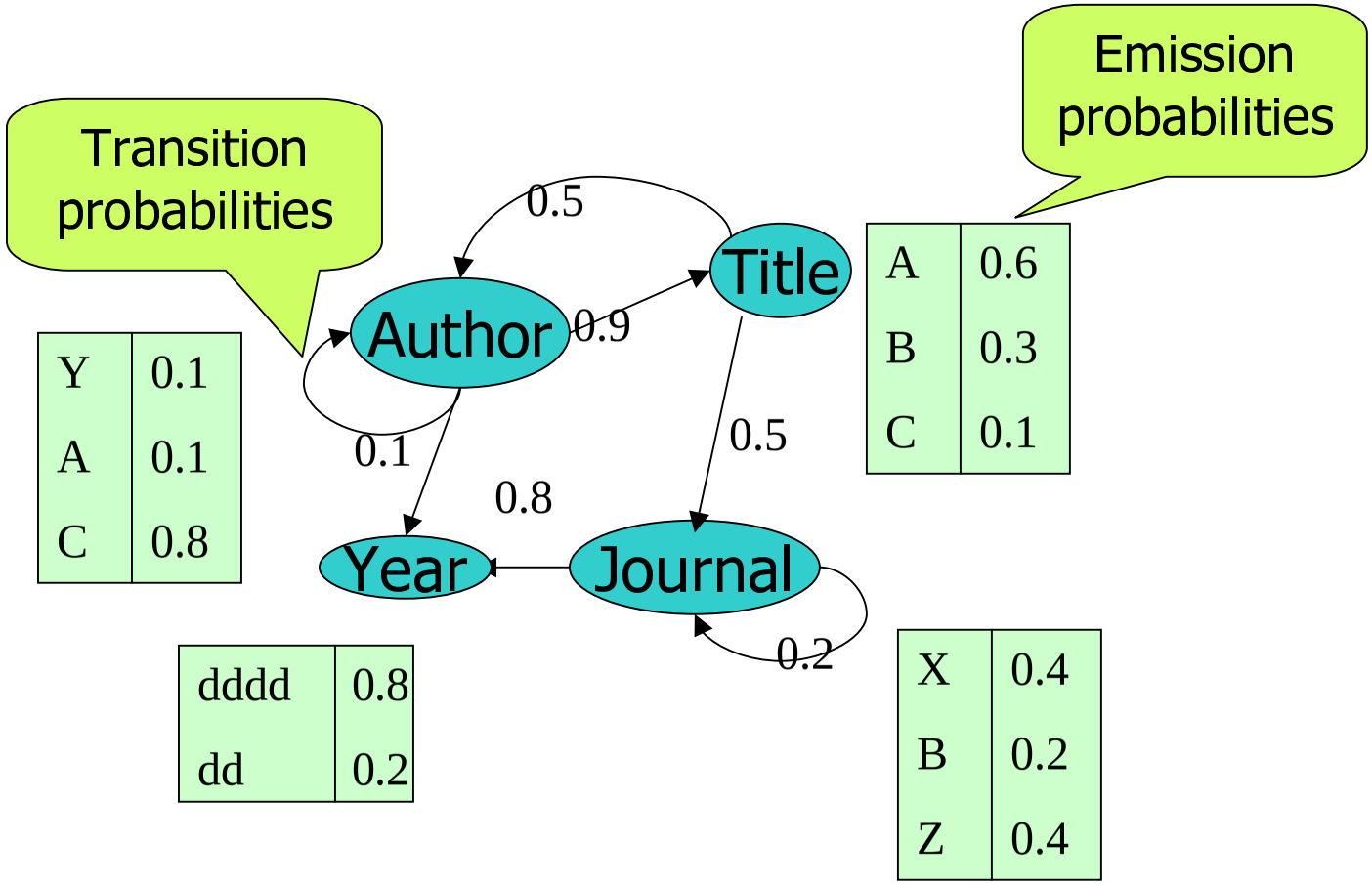


When more than one rule fires apply more specific rule

All allowable permutations must appear in training data

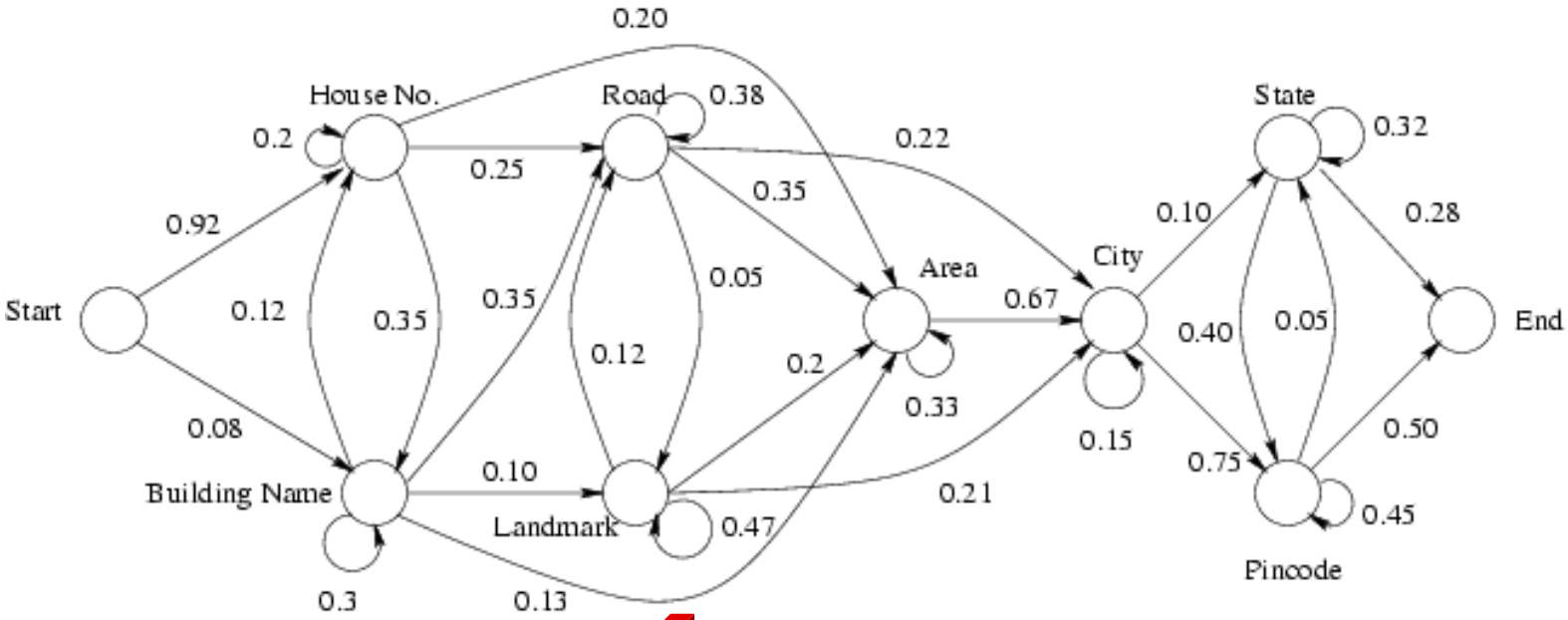
IE with Hidden Markov Models

- Probabilistic models for IE



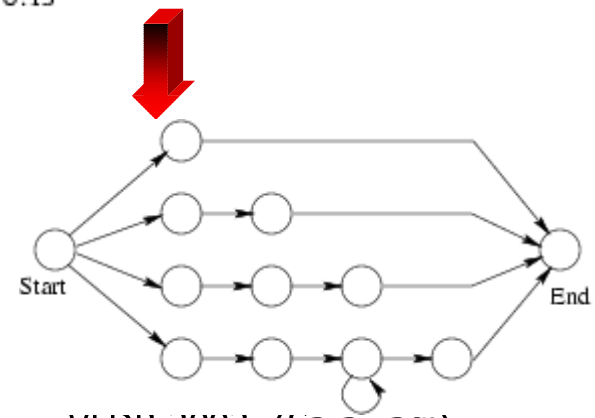
HMM Structure

- Naïve Model: One state per element



- Nested model

Each element another HMM



HMM Dictionary

- For each word (=feature), associate the probability of emitting that word
  - Multinomial model
- More advanced models with overlapping features of a word,
  - example,
    - part of speech,
    - capitalized or not
    - type: number, letter, word etc
  - Maximum entropy models (McCallum 2000)

Learning model parameters

- When training data defines unique path through HMM
  - Transition probabilities
    - Probability of transitioning from state i to state j =

$$\frac{\text{number of transitions from i to j}}{\text{total transitions from state i}}$$

  - Emission probabilities
    - Probability of emitting symbol k from state i =

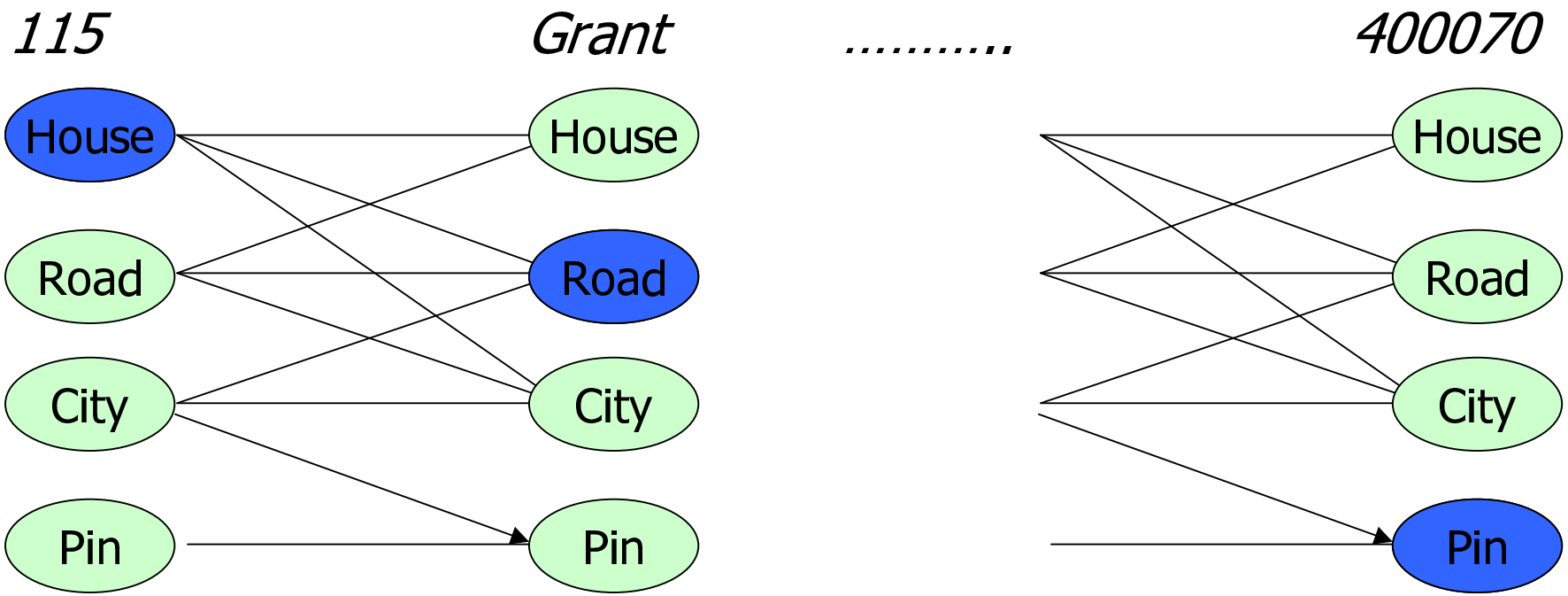
$$\frac{\text{number of times k generated from i}}{\text{number of transition from I}}$$

- When training data defines multiple path:
  - A more general EM like algorithm (Baum-Welch)

Using the HMM to segment

- Find highest probability path through the HMM.
- Viterbi: quadratic dynamic programming algorithm

115 Grant street Mumbai 400070

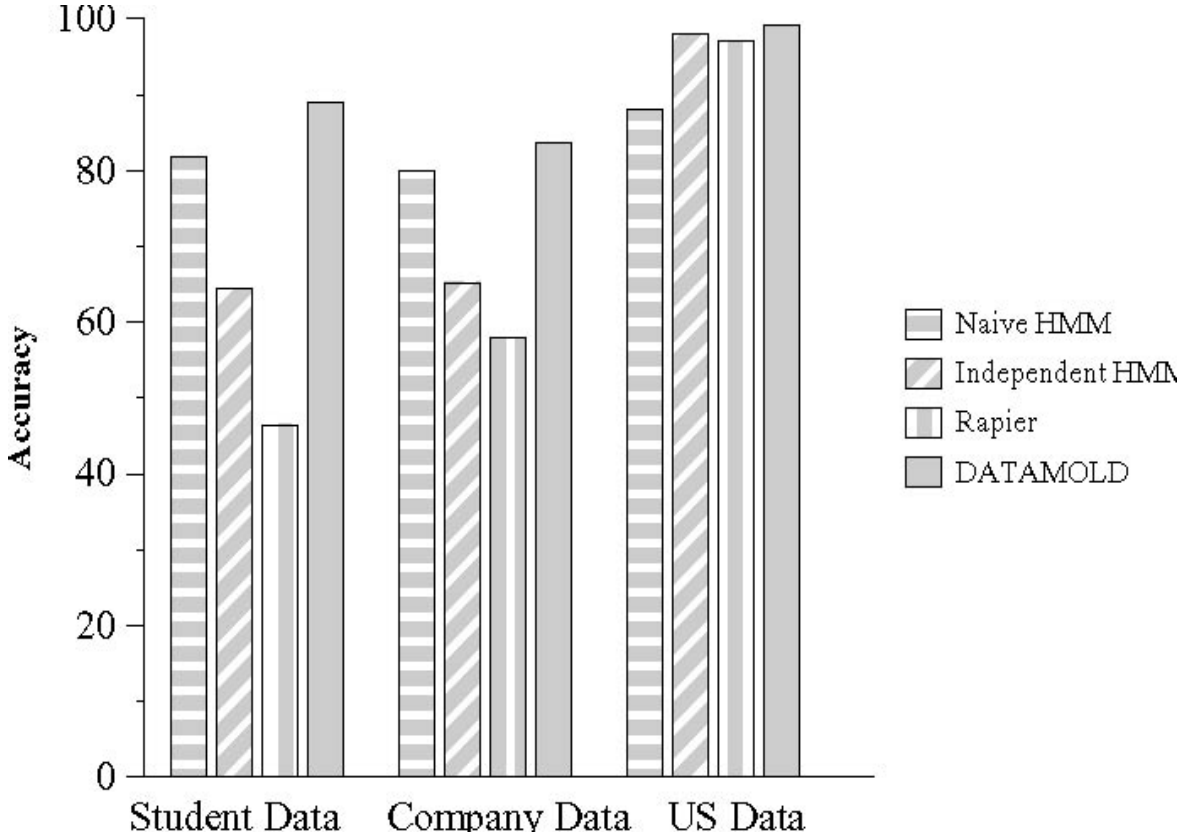


Comparative Evaluation

- 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)

HMM approach: summary

Inter-element sequencing	⇨ Outer HMM transitions
Intra-element sequencing	⇨ Inner HMM
Element length	⇨ Multi-state Inner HMM
Characteristic words	⇨ Dictionary
Non-overlapping tags	⇨ Global optimization

Information Extraction: summary

Feature engineering is key: have to model how to combine them without undue complexity

Rule-based

- And/or combination with heuristics to control firing
- Brittle to variations in data
- Require lesser training data, wrappers reported to learn with < 10 examples
- Used in HTML wrappers

Probabilistic

⇨ Joint probability distribution,more elegant

⇨ Might get hard in general

⇨ Can handle variations

⇨ Used for text segmentation and NE extraction

Outline

- Information Extraction
  - Rule-based methods
  - Probabilistic methods
- Duplicate elimination
- Reducing the need for training data:
  - Active learning
  - Bootstrapping from structured databases
  - Semi-supervised learning
- Summary and research problems

The de-duplication problem

Given a list of semi-structured records,
find all records that refer to a same entity

- Example applications:
  - Data warehousing: merging name/address lists
    - Entity:
      a) Person
      b) Household
  - Automatic citation databases (Citeseer): references
    - Entity: paper

De-duplication:

- is not unsupervised clustering
- precise external notion of correctness

Challenges

- Errors and inconsistencies in data
- Spotting duplicates might be hard as they may be spread far apart:
  - may not be group-able using obvious keys
- Domain-specific
  - Existing manual approaches require re-tuning with every new domain

Example: citations from Citeseer

- Our prior:
  - duplicate when author, title, booktitle and year match..

- Author match could be hard:
  - `L. Breiman, L. Friedman, and P. Stone, (1984).`
  - `Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone.`

- Conference match could be harder:
  - `In VLDB-94`
  - `In Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, September 1994.`

- Fields may not be segmented,
- Word overlap could be misleading

Non-duplicates with lots of word overlap

- H. Balakrishnan, S. Seshan, and R. H. Katz., Improving Reliable Transport and Hando Performance in Cellular Wireless Networks, ACM Wireless Networks, 1(4), December 1995.
- H. Balakrishnan, S. Seshan, E. Amir, R. H. Katz, "Improving TCP/IP Performance over Wireless Networks," Proc. 1st ACM Conf. on Mobile Computing and Networking, November 1995.

Duplicates with little overlap even in title

- Johnson Laird, Philip N. (1983). Mental models. Cambridge, Mass.: Harvard University Press.
- P. N. Johnson-Laird. Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Cambridge University Press, 1983

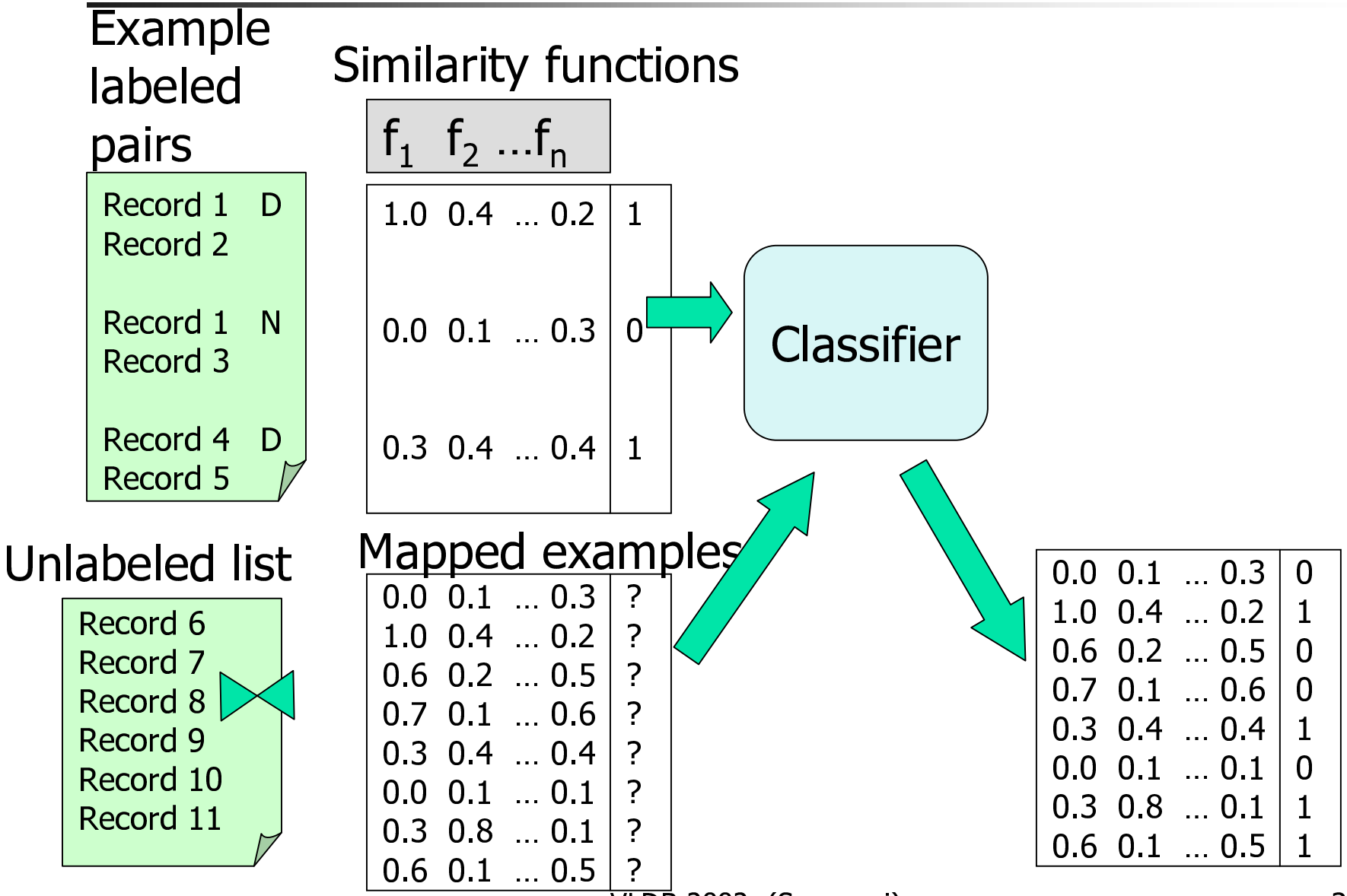
Learning the de-duplication function

Given examples of duplicates and non-duplicate pairs, learn to predict if pair is duplicate or not.

Input features:

- Various kinds of similarity functions between attributes
  - Edit distance, Soundex, N-grams on text attributes
  - Absolute difference on numeric attributes
- Some attribute similarity functions are incompletely specified
  - Example: weighted distances with parameterized weights
  - Need to learn the weights first.

The learning approach



Learning attribute similarity functions

- String edit distance with parameters:
  - C(x,y): cost of replacing x with y
  - d: cost of deleting a character
  - i: cost of inserting a character
- Learning parameters from examples showing matchings
- Transformed Examples:
  - sequence of
    - Match
    - Insert
    - Deletes
- [Bilenko & Mooney, 2002]
- [Ristad & Yianilos 1998
  - Train a stochastic model on sequence

Akme	Inc.
Acme	Incorporated
Mrmm	mmmidddd

Summary: De-deduplication

- Previous work concentrated on designing good static, domain-specific string similarity functions
- Recent spate of work on dynamic learning-based approach appears promising
- Two levels:
  - Attribute-level: Tuning parameters of existing string similarity functions to match examples
  - Record-level: Classifiers like SVMs and decision trees used to combine the similarity along various attributes saving the effort of tuning thresholds and conditions

Outline

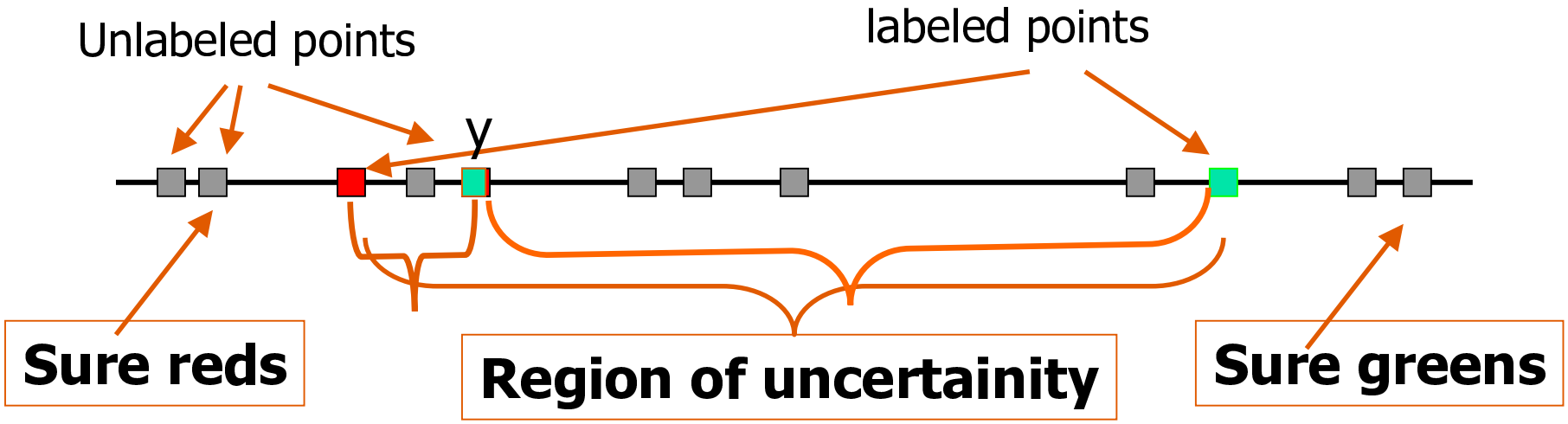
- Information Extraction
  - Rule-based methods
  - Probabilistic methods
- Duplicate elimination
- Reducing the need for training data:
  - Active learning
  - Bootstrapping from structured databases
  - Semi-supervised learning
- Summary and research problems

Active learning

- Ordinary learner:
  - learns from a fixed set of labeled training data
- Active learner:
  - Selects unlabeled examples from a large pool and interactively seeks their labels from a user
  - Careful selection of examples could lead to faster convergence
  - Useful when unlabeled examples are abundant and labeling them requires human effort

Example: active learning

Assume: Points from two classes (red and green) on a real line perfectly separable by a single point separator



Need greatest expected reduction in the size of the uncertainty region → That often corresponds to point with highest prediction uncertainty

Measuring prediction certainty

- Classifier-specific methods
  - Support vector machines:
    - Distance from separator
  - Naïve Bayes classifier:
    - Posterior probability of winning class
  - Decision tree classifier:
    - Weighted sum of distance from different boundaries, error of the leaf, depth of the leaf, etc
- Committee-based approach:

  (Seung, Opper, and Sompolinsky 1992)
  - Disagreements amongst members of a committee
  - Most successfully used method

Forming a classifier committee

Randomly perturb learnt parameters

- Probabilistic classifiers:.
  - Sample from posterior distribution on parameters given training data.
  - Example: binomial parameter p has a beta distribution with mean p
- Discriminative classifiers:
  - Random boundary in uncertainty region

Committee-based algorithm

- Train k classifiers C1, C2,.. Ck on training data
- For each unlabeled instance x
  - Find prediction y1,.., yk from the k classifiers
  - Compute uncertainty U(x) as entropy of above y-s
- ~~Pick instance with highest uncertainty~~
- Sampling for representativeness:
  - With weight as U(x), do weighted sampling to select an instance for labeling.

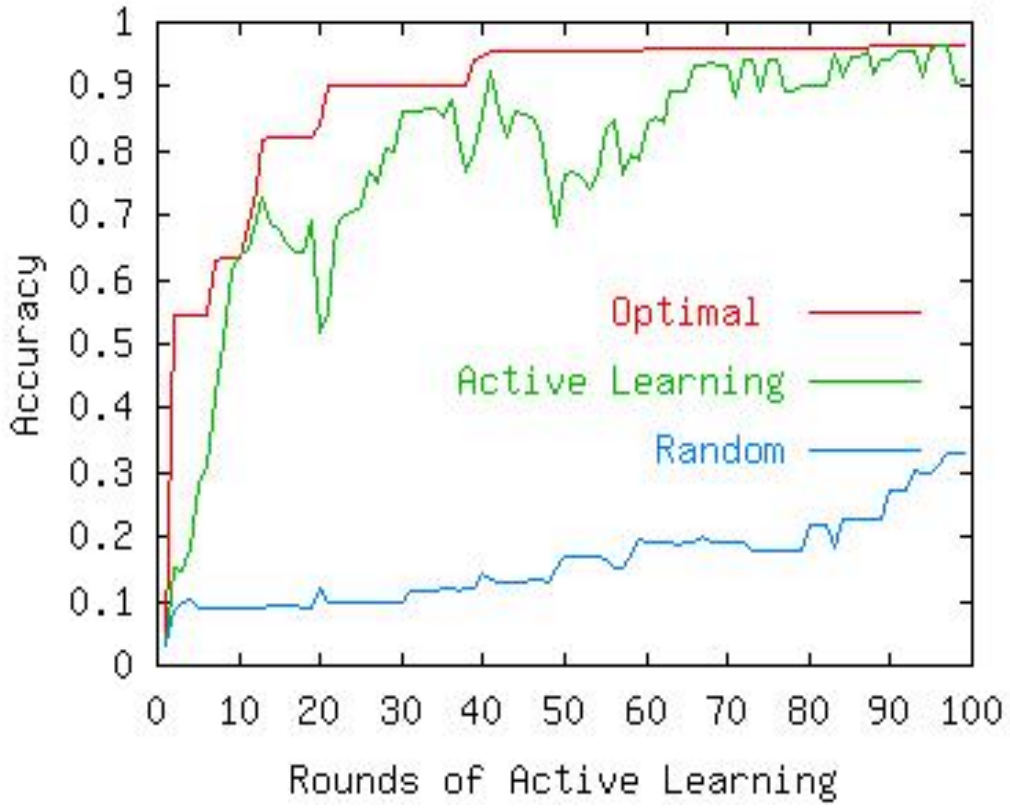
Active learning in deduplication with decision trees

Forming committee of trees by random perturbation

- Selecting split attribute
  - Normally: attribute with lowest entropy
  - Perturbed: random attribute within close range of lowest
- Selecting a split point
  - Normally: midpoint of range with lowest entropy
  - Perturbed: a random point anywhere in the range with lowest entropy

Speed of convergence

Learning deduplication function on Bibtex entries



- With 100 pairs:
  - Active learning: 97% (peak)
  - Random: only 30%

(from Sarawagi 2002)

Active learning in IE with HMM

Forming committee of HMMs by random perturbation

- Emission and transition probabilities are independent multinomial distributions.
- Posterior distribution for Multinomial parameters:
  - Dirichlet with mean estimated as using maximum likelihood
- Results on part of speech tagging (Dagan 1999)
  - 92.6% accuracy using active learning with 20,000 instances as against 100,000 random

Active learning in rule-based IE

Stalker (Muslea et al 2000)

- Learn two classifiers:
  - one based on a forward traversal of the document,
  - second based on a backward traversal
- Select for labeling those records that get conflicting prediction from the two
- Performance: 85% accuracy without active learning yield 94% with active learning

Bootstrapping from structured databases

- Given a database of structured elements
  - Example: collection of structured bibtex entries
- Segment to best match with the database
- HMM:
  - Initialize dictionary using database
  - Learn transitions using Baum Welch on unlabeled data
  - Assigning probabilities hard
  - Still open to investigation
- Rule-based IE: (Snowball, Agichtein 2000)

Semi-supervised learning

- Can unlabeled data improve classifier accuracy?
- Possibly, for probabilistic classifiers like HMMs
  - Use labeled data to train an initial model
  - Use Baum Welch on unlabeled data to refine model to maximize data likelihood
  - Unfortunately, no gain in accuracy reported (Seymore 1999)
- Needs further investigation

Unsupervised learning in duplicate elimination

[Tailor 2002]

- Cluster the similarity vectors of record pairs into three groups
- Label the clusters based on distance to the ideal duplicate and non-duplicate vector.
- (optional) Train classifier on this labeled data

- Results: 79.8% accuracy on Walmart's items table.

Summary

- Information Extraction
  - Various levels of complexity depending on input
    - Segmentation, HTML wrappers, free-format
  - Model-type:
    - Rule-based and probabilistic (HMM)
    - Independent or simultaneous
  - Several research prototypes in each type
- Duplicate elimination
  - Challenging because of variations in data format
  - Learning applied to design deduplication function

Manual Vs learning approach

Manual

- Inspect patterns
- Code scripts
- Requires high-skill programmer

Learning

- Label examples
- Choose & train model
- Low-skill, cheaper labor for most part
- Feature design and model selection requires very high skill

Summary

- Reducing need for labeled data
  - Active learning
    - Various methods proposed
    - Committee-based sampling most popular
    - Application with
      - HMM for IE
      - Decision trees for deduplication

Topics of further research

- Information Extraction:
  - Exploiting higher-level structures in input data, e.g. trees, tables
  - Integrated learning in the presence of a large structured DB, small labeled data and large unlabeled data
  - Efficiency in the presence of a large database/dictionary
  - Wrappers at the website level involving several structured tables
- Duplicate elimination
  - Multi-table de-duplication
  - Integrating semi-supervised and active learning
  - Efficient active learning without requiring materialization of all possible pairs
  - Efficient evaluation of a de-duplication function

Topics of further research

- Combining machine learning of extraction patterns with human generated scripts
- Updating models as data arrives: continuous learning
- Going from research prototypes to robust products and toolkits

References

- **General**
  - H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C. Saita. Declarative data cleaning: Language, model and algorithms. VLDB, 2001.
  - S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. IEEE Computer, 32(6):67-71, 1999.
  - A. McCallum, K. Nigam, J. Reed, J. Rennie, and K. Seymore. Cora: Computer science research paper search engine. http://cora.whizbang.com/, 2000.
  - IEEE Data Engineering special issue on Data Cleaning. http://www.research.microsoft.com/research/db/debull/A00dec/issue. htm, December 2000.
  - M. A. Hernandez and S. J. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery, 2(1), 1998.
- **Information extraction**
  - E. Agichtein, L. Gravano, “Snowball: Extracting relations from large plaintext collections", ACM Intl. Conf. on Digital Libraries“ 2000
  - D M. Bikel, S Miller, R Schwartz and R. Weischedel, "Nymble: a high-performance learning name-finder", ANLP 1997,
  - Vinayak R. Borkar, Kaustubh Deshmukh, and Sunita Sarawagi. Automatic text segmentation for extracting structured records. SIGMOD 2001.
  - Mary Elaine Calif and R. J. Mooney. Relational learning of pattern-match rules for information extraction. AAAI 1999.
  - D Freitag and A McCallum, Information Extraction with HMM Structures Learned by Stochastic Optimization, AAAI 2000
  - A. McCallum and D. Freitag and F. Pereira, Maximum entropy Markov models for information extraction and segmentation, ICML-2000

References

- K Seymore, A McCallum, R Rosenfeld. Learning Hidden Markov Model structure for information extraction. AAAI Workshop on Machine Learning for Information Extraction, 1999.
- S. Soderland. Learning information extraction rules for semi-structured and free text. Machine Learning, 34, 1999.

- **Wrappers**
  - C.Y. Chung, M. Gertz, and N. Sundaresan. Reverse engineering for web data: From
  - visual to semantic structures. ICDE 2002.
  - William W. Cohen, Matthew Hurst, and Lee S. Jensen. A exible learning system for wrapping tables and lists in html documents. WWW 2002.
  - David W. Embley, Y. S. Jiang, and Yiu-Kai Ng. Record-boundary discovery in web documents. In SIGMOD 1999.
  - C.-N. Hsu and M.-T. Dung. Generating finite-state transducers for semistructured data extraction from the web. Information Systems Special Issue on Semistructured Data, 23(8), 1998.
  - N. Kushmerick, D.S. Weld, and R. Doorenbos. Wrapper induction for information extraction. IJCAI, 1997.
  - L. Liu, C. Pu, and W. Han. Xwrap: An XML-enabled wrapper construction system for web information sources. ICDE, 2000.
  - Ion Muslea, Steven Minton and Craig A. Knoblock, Hierarchical Wrapper Induction for Semistructured Information Sources, "Autonomous Agents and Multi-Agent Systems", 2001.
  - Jussi Myllymaki. Effective web data extraction with standard XML technologies. WWW, 2001.

References

- **Duplicate elimination**
  - A Z. Broder, S C. Glassman, M S. Manasse, Geoffrey Zweig, “Syntactic Clustering of the Web”, WWW 1997
  - M. G. Elfeky, V. S. Verykios, A.K. Elmagarmid, “Tailor: A record linkage toolkit”, ICDE 2002.
  - S Sarawagi and Anuradha Bhamidipaty, Interactive deduplication using active learning, ACM SIGKDD 2002
  - W. E. Winkler. Matching and record linkage. In B. G. C. et al, editor, Business Survey Methods, pages 355-384. New York: J. Wiley, 1995.
- **Active and semi-supervised learning**
  - Shlomo Argamon-Engelson and Ido Dagan. Committee-based sample selection for probabilistic classififers. J. of Artificial Intelligence Research, 11:335--360, 1999.
  - Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. Machine Learning, 28(2-3):133-168, 1997.
  - Ion Muslea, Steve Minton, and Craig Knoblock. “Selective sampling with redundant views". AAAI, 2000
  - H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In Computational Learing Theory, pages 287-294, 1992.
  - T. Zhang and F. J. Oles. A probability analysis on the value of unlabeled data for classification problems. ICML, 2000