Text Learning and Information Extraction

Machine Learning and Text

- Textual data is ubiquitous and ever-important
  - WWW, digital libraries, LexisNexis, Medline, news, ….
- Machine learning is required for high performance on key tasks for textual data
  - Retrieval (search, question answering, extraction)
    - Learn to (accurately) compute relevance between query and documents
  - Classification
    - Learn to (accurately) categorize documents
  - Clustering
    - Learn to (accurately) group documents
  - Object identification
    - Learn to (accurately) determine whether textual strings are equivalent

Text as Data

- Representing documents: a continuum of richness
  - **Vector-space**: text is a |*V*|-dimensional vector (*V* is vocabulary of all possible words), order is ignored ("bag-of-words")
  - **Sequence**: text is a string of contiguous tokens/characters
  - **Language-specific**: text is a sequence of contiguous tokens along with various syntactic, semantic, and pragmatic properties (e.g. part-of-speech features, semantic roles, discourse models)
- Higher representation richness leads to higher computational complexity, more parameters to learn, etc., but may lead to higher accuracy

Representation richness

Natural Language Processing

- An entire field focused on tasks involving syntactic, semantic, and pragmatic *analysis* of natural language text
  - Examples: part-of-speech tagging, semantic role labeling, discourse analysis, text summarization, machine translation.
- Using machine learning methods for automating these tasks is a very active area of research, both for ML and NLP researchers
  - Text-related tasks rely on learning algorithms
  - Text-related tasks present great challenges and research opportunities for machine learning

Information Extraction

- Identify specific pieces of information (data) in a unstructured or semi-structured textual document
- Transform unstructured information in a corpus of documents or web pages into a structured database
- Can be applied to different types of text
  - Newspaper articles, web pages, scientific articles, newsgroup messages, classified ads, medical notes, …
- Can employ output of Natural Language Processing tasks for enriching the text representation ("NLP features")

Sample Job Posting

Subject: US-TN-SOFTWARE PROGRAMMER
Date: 17 Nov 1996 17:37:29 GMT
Organization: Reference.Com Posting Service
Message-ID: <56nigp$mrs@bilbo.reference.com>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com

Sample Job Posting

Subject: US-TN-SOFTWARE PROGRAMMER
Date: 17 Nov 1996 17:37:29 GMT
Organization: Reference.Com Posting Service
Message-ID: <56nigp$mrs@bilbo.reference.com>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based **Voice Mail** systems. Experienced in **C Programming**. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer **5** years or more experience with **PC** Based **Voice Mail**, but will consider as little as **2** years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is **DOS**. May go to **OS-2** or **UNIX** in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com

Extracted Job Template

computer_science_job
id: 56nigp$mrs@bilbo.reference.com
title: SOFTWARE PROGRAMMER
salary:
company:
recruiter:
state: TN
city:
country: US
language: C
platform: PC \ DOS \ OS-2 \ UNIX
application:
area: Voice Mail
req_years_experience: 2
desired_years_experience: 5
req_degree:
desired_degree:
post_date: 17 Nov 1996

Medline Corpus

TI - Two potentially oncogenic cyclins, cyclin A and cyclin D1, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the Rb protein

AB - Originally identified as a ‘mitotic cyclin’, cyclin A exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an S-phase-promoting factor (SPF) as well as a candidate proto-oncogene …

Moreover, cyclin D1 was found to be phosphorylated on tyrosine residues in vivo and, like cyclin A, was readily phosphorylated by pp60c-src in vitro.

In synchronized human osteosarcoma cells, cyclin D1 is induced in early G1 and becomes associated with p9Ckshs1, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing’s sarcoma cells demonstrated that cyclin D1 is associated with both p34cdc2 and p33cdk2, and that cyclin D1 immune complexes exhibit appreciable histone H1 kinase activity …

Medline Corpus

TI - Two potentially oncogenic cyclins, **cyclin A** and **cyclin D1**, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the **Rb** protein

AB - Originally identified as a ‘mitotic cyclin’, **cyclin A** exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an **S-phase-promoting factor** (**SPF**) as well as a candidate proto-oncogene …

Moreover, **cyclin D1** was found to be phosphorylated on tyrosine residues in vivo and, like **cyclin A**, was readily phosphorylated by **pp60c-src** in vitro.

In synchronized human osteosarcoma cells, **cyclin D1** is induced in early G1 and becomes associated with **p9Ckshs1**, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing’s sarcoma cells demonstrated that **cyclin D1** is associated with both **p34cdc2** and **p33cdk2**, and that **cyclin D1** immune complexes exhibit appreciable histone H1 kinase activity …

Medline Corpus: Relation Extraction

TI - Two potentially oncogenic cyclins, **cyclin A** and **cyclin D1**, share common properties of subunit ~~configuration, tyrosine~~ phosphorylation and physical association with the **Rb** protein

AB - Originally identified as a ‘mitotic cyclin’, **cyclin A** exhibits properties of growth factor sensitivity, susceptibility to viral subversion and association with a tumor-suppressor protein, properties which are indicative of an **S-phase-promoting factor** (**SPF**) as well as a candidate proto-oncogene …

Moreover, **cyclin D1** ~~was found~~ to be phosphorylated on tyrosine residues in vivo and, like **cyclin A**, ~~was readily phosphorylated by~~ **pp60c-src** in vitro.

In synchronized human osteosarcoma cells, **cyclin D1** is induced in early G1 and becomes associated with **p9Ckshs1**, a Cdk-binding subunit.

Immunoprecipitation experiments with human osteosarcoma cells and Ewing’s sarcoma cells ~~demonstrated that~~ **cyclin D1** ~~is associated with both~~ **p34cdc2** and **p33cdk2**, and that **cyclin D1** immune complexes exhibit appreciable histone H1 kinase activity …

Named Entity Recognition and Linkage

- Without an underlying database, simply recognizing certain *named entities* can be very useful for better searching and indexing
- Identifying *co-referent* entities is important for performance: instance of *record linkage* problem

The Senate Democratic leader, Harry Reid of Nevada, said Tuesday that he would oppose the confirmation of Judge John G. Roberts Jr. as chief justice, surprising both the White House and fellow Democrats still conflicted about how to vote.

IE History: from MUC to Biology

- 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 has many applications for of particular interest to the intelligence community (CIA, NSA)
- Recently, much interest from .com's and biologists

Other Applications

- Smarter email
  - GMail shows links for address maps and tracking UPS packages
- Web classifieds and internet shopping
  - Craigslist aggregators, Froogle
- Job postings
  - Newsgroups (Rapier from austin.jobs), Web pages: Flipdog
- Job resumes
  - BurningGlass, Mohomine
- Seminar announcements
- Company/university information from the web
- Apartment rental ads
- Molecular biology information from MEDLINE

IE via Extraction Patterns

- In many domains, documents are *semi-structured:* text has regularities that allow hand-constructing extraction rules for selecting fields of interest
- Example: extracting book pages from amazon.com

```
<b class="sans">The Dream Machine : J.C.R. Licklider and the Revolution That Made
Computing Personal (Paperback)</b><br>
<font face=verdana,arial,helvetica size=-1>
by <a href="/exec/obidos/search-handle-url/index=books&field-author-
exact=M.%20Mitchell%20Waldrop/002-8262876-4304052">M. Mitchell Waldrop</a><br>
</font> <br> <span class="small"> <span class="small">
<b>List Price:</b> <span class=listprice>$16.00</span><br>
<b>Our Price: <font color=#990000>$10.88</font></b><br>
```

```
<b>List Price:</b> <span class=listprice>$22.00</span><br>
<b>Our Price: <font color=#990000>$14.96</font></b><br>
```

```
<b>List Price:</b> <span class=listprice>$25.00</span><br>
<b>Our Price: <font color=#990000>$16.50</font></b><br>
```

Simple Extraction Patterns

- Specify an item to extract for a slot using a regular expression pattern.
  - Price pattern: "\b\$\d+(\.\d{2})?\b"
- May require preceding (pre-filler) pattern to identify proper context.
  - Amazon list price:
    - Pre-filler pattern: "List Price: "
    - Filler pattern: "\$\d+(\.\d{2})?\b"
- May require succeeding (post-filler) pattern to identify the end of the filler.
  - Amazon list price:
    - Pre-filler pattern: "List Price: "
    - Filler pattern: ".+"
    - Post-filler pattern: ""

Web Extraction

- Web pages are often generated automatically based on an underlying database
- An IE system for such generated pages allows the web site to be viewed as a structured database
- An extractor for a semi-structured web site is sometimes referred to as a *wrapper*
- While manual pattern construction may be easy, the problem of *wrapper maintenance* arises with website changes
- Wrapper adaptation is similar to the problem of extraction from unstructured text: the task is to *learn* a wrapper given *training examples*

Learning for IE

- Given examples of labeled text, learn how to label tokens (or groups of tokens).
- Basic approach: token classification
  - Treat each token as an isolated instance to be classified.
  - Features include token word, neighbors, capitalization, …
  - Use labeled data as a training set: fields to extract are positive examples, other tokens are negative examples
- Example: biomedical text, protein name extraction

O	*O*	*O*	*O*	*O*	*I*	*O*	*I*	*I*	*O*	*O*	…
to	map	the	interaction	of	PTHrP	with	importin	beta	using	a	…

IE via Token Classification (1)

to map the interaction of PTHrP with importin beta using a …

x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_8 x_9 x_{10} …

- Each token is represented by a feature vector
- Possible features: token value, is_dictionary_word, has_uppercase, ends_with_"**-in**", is_noun

x_1={'to', T, F, F, F} $y_1=O$
x_2={'map', T, F, F, F} $y_2=O$
…
x_6={'PTHrP', F, T, F, T} $y_6=I$
x_7={'with', T, F, F, F} $y_7=O$
x_8={'importin', F, F, T, T} $y_8=I$
x_9={'beta', T, F, F, T} $y_9=I$
…

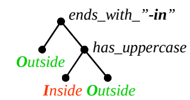
Task: given training data, learn a classifier that labels every new t_i as either *Inside* or *Outside*:

x_i={''Cyclin", F, T, T, T}

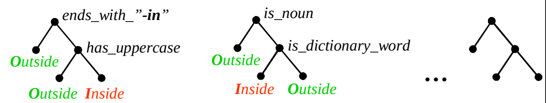
y_i=***?***

IE via Token Classification – Example 1

- Decision Trees



- Ensembles of decision trees …
  - Multiple models (e.g., Boosting, Bagging)



IE via Token Classification – Example 2

- Naïve Bayes classifier: assuming features are *independent*, find probability of class using Bayes theorem:

$$p(y_i|x_i) = \frac{p(y_i)\ p(x_i|y_i)}{p(x_i)}$$

  - Feature independence means that for each x_i:

$$p(x_i) = p(x_i = \{x_{i1}, \ldots, x_{ik}\}) = \prod_{i=1}^{K} p(x_{ik})$$
$$p(x_i|y_i) = \prod_{i=1}^{K} p(x_{ik}|y_i)$$

  - For individual features, probabilities can be obtained from training data frequencies:
    - $p(y_i=I) = 0.1$
    - $p(x_{i1}=\text{'PTHrP'} \mid y_i=I) = 0.9$, $p(x_{i1}=\text{'beta'} \mid y_i=I) = 0.7$, …
    - $p(x_{i2}=T \mid y_i=I) = 0.4$, $p(x_{i3}=T \mid y_i=I) = 1.0$, …
  - Probability of labeling is computed for each token:

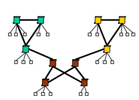
$$p(y_i = I | x_i = \{x_{i1}, \ldots, x_{i5}\}) = \frac{p(y_i=I)\ \prod_{k=1}^{5} p(x_{ik}|y_i=I)}{\prod_{k=1}^{5} p(x_{ik})}$$

IE via Token Classification: Shortcomings of the Single-token approach

- Natural language has very rich structure (syntax, semantics, topical structure, …)
  - Many dependencies exist between words within the sentence
  - "Myopic" classification that considers one token a time is ignoring the dependencies
- In many IE tasks, fields of interest are composed of *several adjacent tokens* ("Harry Reid", "cyclin D1")
  - Labels of adjacent tokens are *related*, labeling decisions should be made *collectively*

Relational Learning and Graphical Models

- Collective, or *relational* learning: instances cannot be treated as *independent*, dependencies between them must be considered
- ***Graphical models*** provide an intuitive and principled framework
  - Instances are nodes, features are attributes of nodes
  - Edges encode dependencies between instance labels and features
  - Example: web page classification

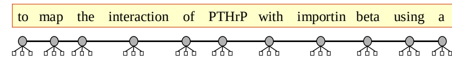


  - Mathematical semantics (probability-based or optimization-based) can be formulated for the graph, leading to a clean problem formulation.
  - Key tasks for any graphical model: (1) ***learning*** (2) ***inference*** (labeling)
  - How do we represent dependency structure for Information Extraction?

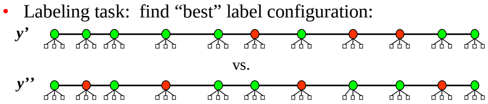
IE and Relational Learning

- Strongest dependencies in text are between adjacent words

to map the interaction of PTHrP with importin beta using a …



- Labeling task: find "best" label configuration:



- What defines "best" configurations?
  - Need to select a mathematical formulation, from which to derive algorithms for learning and for inference

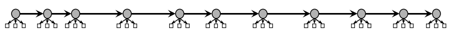
IE and Relational Learning – HMMs (1)

- Hidden Markov Models (HMMs): a ***generative*** model
- Assumes data is produced by a generative process

  For each word: 

  1. Given the label of the current word, generate a label for next word from a distribution (roll a "next label die" for current label value)
  2. Given next word's label, generate its features from a feature distribution (roll a "feature die" for next label value from previous step)
- If natural text was generated like this…

to	map	the	interaction	of	PTHrP	with	importin	beta	using	a	…



IE and Relational Learning – HMMs (2)

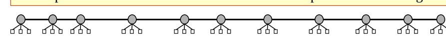
- Learning HMM parameters: Baum-Welch algorithm
  - Parameters are "next label" and "feature" die probabilities
  - E.g. $p(y_i=I \mid y_{i-1} = O)$, $p(x_{ji}=\text{"PTHrP"} \mid y_i=I)$, …
  - "Optimal" probabilities maximize likelihood of observed training data
- Inference in HMMs: Viterbi algorithm
  - Most likely label configuration can be computed in linear time
- Shortcomings of HMMs
  - No easy way to include overlapping, co-dependent features (x_{jk} is "has_uppercase AND NOT is_in_dictionary"): generative model cannot have arbitrary features due to probabilistic semantics
  - "Label bias problem": during inference (and training) decisions are made *locally:* no way to trade off decisions at different positions against each other.
- All details on HMMs coming a few weeks!

IE and Relational Learning – CRFs (1)

- Conditional Random Fields (CRFs) overcome the problems of HMMs
  - A *generative* model is replaced by *the log-linear discriminative model* for both features and inter-label dependencies; this allows arbitrary (possibly overlapping) features
  - Label bias problem is solved by normalizing probabilities of labels for the entire sequence, leading to better performance
- CRFs are *undirectional* graphical models
- Bad news: no simple "dice-driven" semantics, difficult learning
- Good news: a much richer, stronger model

IE and Relational Learning – CRFs (2)

to	map	the	interaction	of	PTHrP	with	importin	beta	using	a	…



- The log-linear model decomposes the conditional probability of each label through an *exponentiated sum of weighted features*
- The overall probability is normalized jointly over *the entire label sequence*

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\exp\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\lambda_k f_k(y_{i-1}, y_i, \mathbf{x}, i)\right)$$

- Learning is difficult: methods find optimal $\boldsymbol{\lambda}$ values by gradient-based search for values that *maximize likelihood of training data*
- Inference can still be performed using Viterbi algorithm

Conclusions

- Textual data
  - Provides many important challenges for machine learning that inspire new research in representation, models, algorithms
  - Presents many applications for evaluating algorithms
- Information extraction
  - A machine learning task that bridges classification, relational learning, and natural language processing
  - High performance requires complicated models and algorithms, employs many features ($O(10^4)$ and higher)
  - Active area of current research: richness of natural text and many applications leave much space for new ideas

Interested in learning more?

- WekaIE: add-on to Weka in the making, email mbilenko@cs.utexas.edu for latest updates
- Several project suggestions are related to IE and record linkage– ask or email mbilenko@cs.utexas.edu