CS 546
Machine Learning in NLP

Structured Prediction:
Theories and Applications
to
Natural Language Processing

Dan Roth
Department of Computer Science
University of Illinois at Urbana-Champaign

- What’s the class about
- How I plan to teach it
- Requirements
- Questions?
Comprehension

A process that maintains and updates a collection of propositions about the state of affairs.

(ENGLAND, June, 1989) - Christopher Robin is alive and well. He lives in England. He is the same person that you read about in the book, Winnie the Pooh. As a boy, Chris lived in a pretty home called Cotchfield Farm. When Chris was three years old, his father wrote a poem about him. The poem was printed in a magazine for others to read. Mr. Robin then wrote a book. He made up a fairy tale land where Chris lived. His friends were animals. There was a bear called Winnie the Pooh. There was also an owl and a young pig, called a piglet. All the animals were stuffed toys that Chris owned. Mr. Robin made them come to life with his words. The places in the story were all near Cotchfield Farm. Winnie the Pooh was written in 1925. Children still love to read about Christopher Robin and his animal friends. Most people don’t know he is a real person who is grown now. He has written two books of his own. They tell what it is like to be famous.

1. Christopher Robin was born in England.
2. Winnie the Pooh is a title of a book.
3. Christopher Robin’s dad was a magician.
4. Christopher Robin must be at least 65 now.

This is an Inference Problem
Why is it difficult?

Meaning

Variability

Language

Ambiguity
Context Sensitive Paraphrasing

- He used a Phillips head to **tighten** the screw.

- The bank owner **tightened** security after a spat of local crimes.

- The Federal Reserve will aggressively **tighten** monetary policy.
Beyond Single Variable Decisions

Example: Relation Extraction: “Works_for”

Jim Carpenter works for the U.S. Government.

The American government employed Jim Carpenter.

Jim Carpenter was fired by the US Government.

Jim Carpenter worked in a number of important positions. .... As a press liaison for the IRS, he made contacts in the white house.

Top Russian interior minister Yevgeny Topolov met yesterday with his US counterpart, Jim Carpenter.

Former US Secretary of Defence Jim Carpenter spoke today...
Textual Entailment

A key problem in natural language understanding is to abstract over the inherent syntactic and semantic variability in natural language.

Is it true that...

Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc. last year.

Yahoo acquired Overture
Overture is a search company
Google is a search company
Google owns Overture

............
Learning

- The process of Abstraction has to be driven by statistical learning methods.

- Over the last decade or so it became clear that machine learning methods are necessary in order to support this process of abstraction.

- But—
Classification: Ambiguity Resolution

Illinois’ bored of education [board]
Nissan Car and truck plant; plant and animal kingdom
(This Art) (can N) (will MD) (rust V) V,N,N
The dog bit the kid. He was taken to a veterinarian; a hospital
Tiger was in Washington for the PGA Tour

→ Finance; Banking; World News; Sports

Important or not important; love or hate
The goal is to learn a function $f: X \rightarrow Y$ that maps observations in a domain to one of several categories.

**Task:** Decide which of \{board, bored\} is more likely in the given context:

- **X:** some representation of: The Illinois’ _______ of education met yesterday...
- **Y:** \{board, bored\}

**Typical learning protocol:**
- Observe a collection of labeled examples $(x,y) \in X \times Y$
- Use it to learn a function $f:X \rightarrow Y$ that is **consistent** with the observed examples, and (hopefully) performs well on new, previously unobserved examples.
Classification is Well Understood

- Theoretically: generalization bounds
  - How many example does one need to see in order to guarantee good behavior on previously unobserved examples.

- Algorithmically: good learning algorithms for linear representations.
  - Can deal with very high dimensionality ($10^6$ features)
  - Very efficient in terms of computation and # of examples. On-line.

- Key issues remaining:
  - Learning protocols: how to minimize interaction (supervision); how to map domain/task information to supervision; semi-supervised learning; active learning; ranking.
  - What are the features? No good theoretical understanding here.

- Is it sufficient for making progress in NLP?
Coherency in Semantic Role Labeling

Predicate-arguments generated should be consistent across phenomena

The touchdown **scored** by Cutler **cemented** the **victory of** the Bears.

<table>
<thead>
<tr>
<th>Verb</th>
<th>Nominalization</th>
<th>Preposition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Predicate:</strong> score</td>
<td><strong>Predicate:</strong> win</td>
<td><strong>Sense:</strong> 11(6)</td>
</tr>
<tr>
<td><strong>A0:</strong> Cutler (scorer)</td>
<td><strong>A0:</strong> the Bears (winner)</td>
<td>“the object of the preposition is the object of the underlying verb of the nominalization”</td>
</tr>
<tr>
<td><strong>A1:</strong> The touchdown (points scored)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Linguistic Constraints:

- A0: the Bears ⇔ Sense(of): 11(6)
- A0: Cutler ⇔ Sense(by): 1(1)
Semantic Parsing

X: “What is the largest state that borders New York and Maryland?”

Y: \[ \text{largest( state( next_to( state(NY) \text{ AND next_to (state(MD))))))} \]

- Successful interpretation involves multiple decisions
  - What entities appear in the interpretation?
  - “New York” refers to a state or a city?
  - How to compose fragments together?
    - \text{state(next_to())} \text{ }\leq\text{ } next_to(state())
Learning and Inference

Natural Language Decisions are **Structured**
- Global decisions in which several local decisions play a role but there are mutual dependencies on their outcome.

- It is essential to make coherent decisions in a way that takes the interdependencies into account. **Joint, Global Inference.**

- Unlike “standard” classification problems, in most interesting NLP problems there is a need to predict values for multiple interdependent variables.

- These are typically called **Structured Output Problems** – and will be the focus of this class.
Statistics or Linguistics?

- Statistical approaches were very successful in NLP

- But, it has become clear that there is a need to move from strictly Data Driven approaches to Knowledge Driven approaches

- **Knowledge**: Linguistics, Background world knowledge

- How to incorporate Knowledge into Learning & Decision Making?

- In many respects Structured Prediction addresses this question.
  - This also distinguishes it from the “standard” study of probabilistic models.
This Class

- Problems
  - that will motivate us

- Perspectives
  - we’ll develop

- What we’ll do
  - and how
Some Examples

- **Part of Speech Tagging**
  - This is a sequence labeling problem
  - The simplest example where (it seems that) the decision with respect to one word depends on the decision with respect to others.

- **Named Entity Recognition**
  - This is a sequence segmentation problem
  - Not all segmentations are possible
  - There are dependencies among assignments of values to different segments.

- **Semantic Role Labeling**
  - Decisions here build on previous decisions (*Pipeline Process*)
  - Clear constraints among decisions
Semantic Role Labeling

Who did what to whom, when, where, why,...

I left my pearls to my daughter in my will.

\[ I]_{A_0} \text{ left } [\text{my pearls}]_{A_1} \text{ to my daughter} \ [A_2] \text{ in my will]_{AM-LOC}. \]

- **A0** Leaver
- **A1** Things left
- **A2** Benefactor
- **AM-LOC** Location

Overlapping arguments

If A2 is present, A1 must also be present.
Algorithmic Approach

- **Identify argument candidates**
  - Pruning [Xue&Palmer, EMNLP’04]
  - Argument Identifier
    - Binary classification (A-Perc)

- **Classify argument candidates**
  - Argument Classifier
    - Multi-class classification (A-Perc)

- **Inference**
  - Use the estimated probability distribution given by the argument classifier
  - Use structural and linguistic constraints
  - Infer the optimal global output

(candidate arguments)
Inference with General Constraint Structure
Recognizing Entities and Relations

Dole's wife, Elizabeth, is a native of N.C.

\[ \begin{array}{c|c}
\text{other} & 0.05 \\
\hline
\text{per} & 0.85 \\
\text{loc} & 0.10 \\
\end{array} \]

\[ \begin{array}{c|c}
\text{other} & 0.10 \\
\hline
\text{per} & 0.60 \\
\text{loc} & 0.30 \\
\end{array} \]

\[ \begin{array}{c|c}
\text{other} & 0.05 \\
\hline
\text{per} & 0.50 \\
\text{loc} & 0.45 \\
\end{array} \]

How to guide the global inference? Why not learn jointly?

\[ \begin{array}{c|c}
\text{irrelevant} & 0.05 \\
\hline
\text{spouse_of} & 0.45 \\
\text{born_in} & 0.50 \\
\end{array} \]

\[ \begin{array}{c|c}
\text{irrelevant} & 0.10 \\
\hline
\text{spouse_of} & 0.05 \\
\text{born_in} & 0.85 \\
\end{array} \]
A Lot More Problems

- POS/Shallow Parsing/NER
- SRL
- Parsing/Dependency Parsing
- Information Extraction/Relation Extraction
- Co-Reference Resolution
- Transliteration
- Textual Entailment
Computational Issues

The Inference Problem
How to solve/make decisions?

The Learning Problem
How to train the model?

Decouple?
Joint Learning vs. Joint Inference

Difficulty of Annotating Data

Indirect Supervision
Constraints Driven Learning

Semi-supervised Learning
Constraints Driven Learning
Perspective

- Models
  - Generative/Descriminative

- Training
  - Supervised, Semi-Supervised, indirect supervision

- Knowledge
  - Features
  - Models Structure
  - Declarative Information

- Approach
  - Unify treatment
  - Demistify
  - Survey key ML techniques used in Structured NLP
This Course

Structure
- Lectures by me; gradually, more lectures by you; a few external talks

Teaching
- On the board, formal, mathematically rigorous

Assignments
- Projects: Two (group) projects, reports and presentations
- Presenting 1-2 papers
- 4 critical surveys
- ??? Group activities; short work sheets

Expectations
- This is an advanced course. I view my role as guiding you through the material and helping you in your first steps as a researcher. I expect that your participation in class, reading assignments, and presentations will reflect independence, mathematical rigor, and critical thinking.
Encouraging Collaborative Learning

- Dimension I: Model
  - On Line, Perceptron Based
  - Exponential Models (Logistic Regression, HMMs, CRFs)
  - SVMs
  - Constrained Conditional Models (Including bank of Classifiers)
  - (Deep Neural Networks?)

- Dimension II: Tasks
  - Basic sequential and Structured Prediction
  - Training Paradigms (Unsupervised & Semi Supervised; Indirect Supervision)
  - Latent Representations
  - Inference and Approximate Inference
  - Applications

- Note that there is a lot of overlaps, several things can belong to several categories. It will make it more interesting.

Before Friday, please send me your preferences (ranked order 1-3) for the Models you want to study and present.

We will form groups of experts that will present all aspects of the model.

By Next Friday, each model group will discuss potential papers and decide who presents what and in what order.
Current Topic Groups

**EXP Models/CRF**
- Girlea, Cordruta
- Mak, Mun Thye
- Wang, Jingjing
- Wang Xiaolong
- Min, Kerui

**SVM**
- Guo, Ruiqi
- Lee, Soo Min
- Ravula, Aniruch
- Wieting, John
- Ye, Dong
- Shah, Smit S.

**CCM**
- Musa, Ryan
- Spirin, Nikita
- Li, Zhijin
- Sun, Guihua
- Pan, Ziying

**Perceptron**
- Sloan, Joseph
- Li Xun
- Liu Haitao
- Truong, Anh
- Ruichen Wang

- **1st** person in each group will be the group representative
- **Wednesday 3/1**: Meet as Groups and choose papers to present (& who does what). You can choose > 1 paper from each area, but try to cover several topics
- Send me the group’s selection (1 per person + a couple additional papers)