CS 546
Machine Learning in NLP

Structured Prediction: Theories and Applications in Natural Language Processing

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

(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?

Variability

Meaning

Language

Ambiguity
Ambiguity

<table>
<thead>
<tr>
<th>It’s a version of <strong>Chicago</strong> – the standard classic <strong>Macintosh</strong> menu font, with that distinctive thick diagonal in the “N”.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chicago</strong> was used by default for <strong>Mac</strong> menus through <strong>MacOS 7.6</strong>, and <strong>OS 8</strong> was released mid-1997.</td>
</tr>
<tr>
<td><strong>Chicago VIII</strong> was one of the early 70s-era <strong>Chicago</strong> albums to catch my ear, along with <strong>Chicago II</strong>.</td>
</tr>
</tbody>
</table>

---

![Chicago Bears](image1)
![Apple](image2)
![Chicago Album](image3)
![MacOS](image4)
![Chicago VIII Album](image5)
![MacOS 7.6](image6)
![Chicago II Album](image7)
Ambiguity

It’s a version of **Chicago** – the standard classic **Macintosh** menu font, with that distinctive thick diagonal in the “N”. **Chicago** was used by default for **Mac** menus through **MacOS 7.6**, and **OS 8** was released mid-1997.

**Chicago VIII** was one of the early 70s-era **Chicago** albums to catch my ear, along with **Chicago II**.
Variability in Natural Language Expressions

Determine if Jim Carpenter works for the government

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.
Russian interior minister Yevgeny Topolov met yesterday with his US counterpart, Jim Carpenter.
Former US Secretary of Defense Jim Carpenter spoke today...

Needs:
- Understanding Relations, Entities and Semantic Classes
- Acquiring knowledge from external resources; representing knowledge
- Identifying, disambiguating & tracking entities, events, etc.
- Time, quantities, processes...

Standard techniques cannot deal with the variability of expressing meaning nor with the ambiguity of interpretation
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
A key problem in natural language understanding is to abstract over the inherent syntactic and semantic variability in natural language.

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 two decades or so it became clear that machine learning methods are necessary in order to support this process of abstraction.

- But—
<table>
<thead>
<tr>
<th>Classification: Ambiguity Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illinois’ <strong>bored</strong> of education</td>
</tr>
<tr>
<td>[board]</td>
</tr>
<tr>
<td>Nissan Car and truck <strong>plant</strong>;</td>
</tr>
<tr>
<td><strong>plant</strong> and animal kingdom</td>
</tr>
<tr>
<td>(This Art) (can N) (will MD) (rust V)</td>
</tr>
<tr>
<td>V,N,N</td>
</tr>
<tr>
<td>The dog bit the kid. <strong>He</strong> was taken to a <strong>veterinarian</strong>; a hospital</td>
</tr>
<tr>
<td>Tiger was in <strong>Washington</strong> for the PGA Tour</td>
</tr>
</tbody>
</table>

→ 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?
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

\(X:\text{"What is the largest state that borders New York and Maryland?"}\)

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

- 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?
    - state(next_to()) \(\preceq\) 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.
Hard Co-reference Problems

- Requires knowledge Acquisition
  - The bee landed on the flower because it had/wanted pollen.
    - Lexical knowledge
  - John Doe robbed Jim Roy. He was arrested by the police.

- The Subj of “rob” is more likely than the Obj of “rob” to be the Obj of “arrest”

- Requires an inference framework that can make use of this knowledge

- NL interpretation is an inference problem best modelled as a knowledge constrained optimization problem over multiple statistically learned models.
This Class

- Problems
  - that will motivate us

- Perspectives
  - we’ll develop
  - It’s not only the learning algorithm...

- 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.

- **Relation Extraction**
  - Works_for (Jim, US-government) ; co-reference resolution

- **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.

\[ [I]_{A0} \text{ left } [\text{my pearls}]_{A1} \text{ to my daughter} [\text{in my will}]_{A2} \text{ AM-LOC}. \]

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

I left my pearls to my daughter in my will.

- Overlapping arguments
- If A2 is present, A1 must also be present.

How to express the constraints on the decisions? How to “enforce” them?
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 distributions given by the argument classifier
- Use structural and linguistic constraints
- Infer the optimal global output

Use the **pipeline architecture’s simplicity** while maintaining uncertainty: keep probability distributions over decisions & use global inference at decision time.
Inference with General Constraint Structure
Recognizing Entities and Relations

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

How to guide global inference? How to learn?
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?

Decouple?
Joint Learning vs.
Joint Inference

The Learning Problem
How to train the model?

Difficulty of Annotating Data

Indirect Supervision
Constraints Driven Learning

Semi-supervised Learning
Constraints Driven Learning
Perspective

- Models
  - Generative/ Discriminative

- 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
  - (Some) Lectures by me. Mostly, lectures/tutorials by you.
  - Group discussions: of teaching material and projects

- Assignments
  - Projects: A group project; reports and presentations
  - Presentations: group paper presentations/tutorials (1-2 papers each)
  - 4 critical surveys
  - Assignment 0 – Machine Learning Preparation.

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

Questions?

Wednesday 9:30

Wednesday 8:30

Peer grading
Organization of Material

- **Dimension I: Models**
  - On Line, Perceptron Based
  - Exponential Models (Logistic Regression, HMMs, CRFs)
  - SVMs
  - Constrained Conditional Models (ILP based formulations)
  - Neural Networks
  - (Markov Logic Networks?)

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

- **Dimension III: Applications**

There is a lot of overlaps, several things can belong to several categories. It will make it more interesting.
Project Groups

- EXP Models/CRF
- Structured SVM
- Neural Networks
- CCM
- Structured Perceptron
- MLNs

**Project: Entities, Relations, and co-reference**

- Eventual goal: gaining from interactions among tasks.
- It is likely that we will be using the ACE data
- Groups will also be responsible to introducing the task, data preparation and evaluation.
Presentations/Tutorial Groups (Tentative)

- EXP Models/CRF
- Structured SVM
- MLN
- Optimization
- Features

- CCM
- Structured Perceptron
- Neural Networks
- Inference
- Latent Representations
Summary

- Class will be taught as a structured seminar + group project

- Meeting: only on Wednesday
  - 9:30: material presentation
  - 8:30: group discussion & project discussions

- Critical readings of papers

- Peer review

- Attendance is mandatory
  - (some of the early Wednesday morning meetings will be per-group)

- If you want to drop the class, do it quickly – not to affect the projects.

- Web site with all the information will be available shortly.

This Friday:

- You will do a prerequisites Machine Learning Exam (take home).
- Send me email with subject: CS546-ML userid
  - I will send you a pdf (with in ID)
  - You solve it (take up to three hours)
  - (don’t include your name on the solution, only the give ID)
  - When you are done, return the pdf with subject: CS546-ML-Solution userid
  - I will then send it out for grading.