Practical Structured Learning for Natural Language Processing

Hal Daumé III
School of Computing
University of Utah
me@hal3.name

CS Colloquium, Brigham Young University, 14 Sep 2006
What Is Natural Language Processing?

Processing of language to solve a problem

- Machine Translation
- Summarization
- Information Extraction
- Question Answering
- Parsing
- Understanding

Input | Output
--- | ---
Arabic sentence | English sentence
Long documents | Short documents
Document | Database
Sentence | Logical sentence

Corpus-based NLP: Collect example input/output pairs and learn

All have innate structure to varying degrees
What is Machine Learning?

Automatically uncovering structure in data

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Bit (or bits)</td>
</tr>
<tr>
<td>Regression</td>
<td>Real number</td>
</tr>
<tr>
<td>Dimensionality reduction</td>
<td>Shorter vector</td>
</tr>
<tr>
<td>Clustering</td>
<td>Partition</td>
</tr>
<tr>
<td>Sequence labeling</td>
<td>Labels</td>
</tr>
<tr>
<td>Reinforcement learning</td>
<td>Policy</td>
</tr>
<tr>
<td>State space + observations</td>
<td></td>
</tr>
</tbody>
</table>

Input: Vector

Output: Vector, Shorter vector, Policy

Partition: Vectors

Labels: Vectors
NLP versus Machine Learning

Machine Translation
Summarization
Information Extraction
Question Answering
Parsing
Understanding

Arabic sentence
Long documents
Document
Corpus + query
Sentence
Sentence

English sentence
Short documents
Database
Answer
Syntactic tree
Logical sentence

Natural Decomposition

Encoding

Ad Hoc Search
Provably Good

Classification
Regression
Dimensionality reduction
Clustering
Sequence labeling

Vector
Vector
Vector
Vectors
Vectors

Bit (or bits)
Real number
Shorter vector
Partition
Labels

Reinforcement learning-inspired

State space
Policy

Structured Learning for Language
Entity Detection and Tracking

Shakespeare scholarship, lively at the best of times, saw the fur flying yesterday after a German academic claimed to have authenticated not just one but four contemporary images of the playwright - and suggested, to boot, that he had died of cancer.

As the National Portrait Gallery planned to reveal that only one of half a dozen claimed portraits of William Shakespeare can now be considered genuine, Prof Hildegard Hammerschmidt-Hummel said she could prove that there were at least four surviving portraits of the playwright.

Why? Summarization, Machine Translation, etc...
Why is EDT Hard?

Long-range dependencies

Syntax, Semantics and Discourse

World Knowledge

Highly constrained outputs

Shakespeare scholarship, lively at the best of times, saw the fur flying yesterday after a German academic claimed to have authenticated not just one but four contemporary images of the playwright - and suggested, to boot, that he had died of cancer.
How is EDT Typically Attacked?

Shakespeare scholarship, lively at the best of times, saw the fur flying yesterday after a German academic claimed to have authenticated not just one but four contemporary images of the playwright - and suggested, to boot, that he had died of cancer.
Learning in NLP Applications

Model

Features

Learning

Search
Result: Searn (= “Search + Learn”)

- Computationally efficient
- Strong theoretical guarantees
- State-of-the-art performance
- Broadly applicable
- Easy to implement
Talk Outline

- Searn: Search-based Structured Prediction

- Experimental Results:
  - Sequence Labeling
  - Entity Detection and Tracking
  - Automatic Document Summarization

- Discussion and Future Work
Structured Prediction

Formulated as a maximization problem:

\[ y = \arg \max_y \ Y_x f \ x, y; \ \theta \]

Search (and learning) tractable only for very simple \( Y(x) \) and \( f \)
Reduction for Structured Prediction

- Idea: view structured prediction in light of search

Each step here looks like it could be represented as a weighted multi-class problem.

Can we formalize this idea?
Error-Limiting Reductions

- A formal mapping between learning problems

**F** maps examples from A to examples from B

**F**⁻¹ maps a classifier h for B to a classifier that solves A

Error limiting: good performance on B implies good performance on A

[Beygelzimer et al., ICML 2005]
Our Reduction Goal

- Error-bounds get composed
- New techniques for binary classification lead to new techniques for structured prediction
- Any generalization theory for binary classification immediately applies to structured prediction

[1] Beygelzimer et al., ICML 2005

“Searn”

Importance Weighted Classification

“Costing”

[Zadronzy, Langford & Abe, ICDM 2003]

“Weighted All Pairs”

[2] Cost-sensitive Classification

“Costing”

[Zadronzy, Langford & Abe, ICDM 2003]
Two Easy Reductions

Costing

[Zadronzy, Langford & Abe, ICDM 2003]

Weighted-All-Pairs

[Beygelzimer, Dani, Hayes, Langford and Zadrozny, ICML 2005]
A First Attempt

At each correct state:
Train classifier to choose next state

Thm (Kääriäinen): There is a 1st-order binary Markov problem such that a classification error \( \Theta \) implies a Hamming loss of:

\[
\frac{T}{2} - \frac{1 - 1 - 2\epsilon}{4\epsilon}^{T+1} \approx \frac{T}{2}
\]
Reducing Structured Prediction

Key Assumption: *Optimal Policy for training data*

Given: input, true output and state;
Return: best successor state

Weak!
Idea

Optimal Policy

Learned Policy
(Features)
Iterative Algorithm: Searn

Set current policy = optimal policy

Repeat:

Decode using current policy

... after a German academic claimed ...

<table>
<thead>
<tr>
<th>GPE</th>
<th>academic</th>
<th>claimed</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>*</td>
<td>L = 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PER</th>
<th>*</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>*</td>
<td>L = 0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ORG</th>
<th>PER</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>*</td>
<td>L = 0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>*</th>
<th>PER</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>*</td>
<td>L = 0.25</td>
</tr>
</tbody>
</table>

Generate classification problems

Learn new multiclass classifier

Interpolate: \( \text{cur} = \alpha \ast \text{cur} + (1 - \alpha) \ast \text{new} \)
Theorem: For conservative $\mathcal{O}$, after $2T^3 \ln T$ iterations, the loss of the learned policy is bounded as follows:

$$L \ h \leq L \ h_0 + 2T \ln T \ l_{avg} + 1 \ \ln T \ \frac{c_{max}}{T}$$

- Loss of the optimal policy
- Average multiclass classification loss
- Worst case per-step loss
Why Does Searn Work?

A classifier tells us how to search

Impossible to train on full space

Want to train only where we'll wind up

Searn tells us where we'll wind up

§3.7 (Searn: Discussion and Conclusions)
Talk Outline

- Searn: Search-based Structured Prediction

- Experimental Results:
  - Sequence Labeling
  - Entity Detection and Tracking
  - Automatic Document Summarization

- Discussion and Future Work
Proof on Concept: Sequence Labels

Spanish Named Entity Recognition

Handwriting Recognition

Chunking+ Tagging

- SVM
- ISO
- CRF
- Searn
- Searn+data

- LR
- SVM
- M3N
- Searn
- Searn+data

Incremental Perceptron

LaSO

CRF

Searn
Talk Outline

- Searn: Search-based Structured Prediction

- Experimental Results:
  - Sequence Labeling
  - Entity Detection and Tracking
  - Automatic Document Summarization

- Discussion and Future Work
Shakespeare scholarship, lively at the best of times, saw the fur flying yesterday after a German academic claimed to have authenticated not just one but four contemporary images of the playwright - and suggested, to boot, that he had died of cancer.
EDT Features

Shakespeare scholarship, lively at the best of times, saw the fur flying yesterday after a German academic claimed to have authenticated not just one but four contemporary images of the playwright - and suggested, to boot, that he had died of cancer.

- Lexical
- Syntactic
- Semantic
- Gazetteer
- Knowledge-based
EDT Results (ACE 2004 data)

Previous state of the art with extra proprietary data
Feature Contributions

- Lex
- Syn
- Sem
- Gaz
- KB

§ 5.6.3 (EDT: Experimental Results)
Talk Outline

- Searn: Search-based Structured Prediction

- Experimental Results:
  - Sequence Labeling
  - Entity Detection and Tracking
  - Automatic Document Summarization

- Discussion and Future Work
New Task: Summarization

That's perfect!

Argentina was still obsessed with the Falkland Islands even in 1994, 12 years after its defeat in the 74-day war with Britain. The country's overriding foreign policy aim continued to be winning sovereignty over the islands.

The Falkland Islands war, in 1982, was fought between Britain and Argentina.

Standard approach is sentence extraction, but that is often deemed too “coarse” to produce good, very short summaries. We wish to also drop words and phrases => document compression
To make search more tractable, we run an initial round of sentence extraction @ 5x length

Argentina was still obsessed with the Falkland Islands even in 1994, 12 years after its defeat in the 74-day war with Britain. The country's overriding foreign policy aim continued to be winning sovereignty over the islands.

Optimal Policy not analytically available; Approximated with beam search.
Example Output (40 word limit)

**Sentence Extraction + Compression:**
Argentina and Britain announced an agreement, nearly eight years after they fought a 74-day war a populated archipelago off Argentina's coast. Argentina gets out the red carpet, official royal visitor since the end of the Falklands war in 1982.

**Vine Growth:**
Argentina and Britain announced to restore full ties, eight years after they fought a 74-day war over the Falkland islands. Britain invited Argentina's minister Cavallo to London in 1992 in the first official visit since the Falklands war in 1982.

**6.7 (Summarization: Error Analysis)**
- Diplomatic ties restored
- Major cabinet member visits
- Exchanges were in 1992
- War between Britain and Argentina
- Falkland war was in 1982
- Cavallo visited UK
- War was 74-days long
Results

Approx. Optimal Vine Growth
Vine Growth (Searn)
Optimal Extract @100
Extract @100
Baseline

[Daumé III and Marcu, DUC 2005]
Talk Outline

- Searn: Search-based Structured Prediction
- Experimental Results:
  - Sequence Labeling
  - Entity Detection and Tracking
  - Automatic Document Summarization
- Discussion and Future Work
What Can Searn Do?

Solve structured prediction problems...

...efficiently.

...with theoretical guarantees.

...with weak assumptions on structure.

...and outperform the competition.

...with little extra code required.
Thanks!

Questions?