Practical Structured Learning for Natural Language Processing

Hal Daumé III

Information Sciences Institute
University of Southern California

hdaume@isi.edu

Committee:
Daniel Marcu (Adviser), Kevin Knight, Eduard Hovy
Stefan Schaal, Gareth James, Andrew McCallum

What Is Natural Lang ssing?

Processing of language t m

All have innate structure to varying degrees

	Input	Output
Machine Translation	Arabic sentence	English sentence
★ **Summarization**	Long documents	Short documents
★ **Information Extraction**	Document	Database
		tree
Understanding	Sentence	Logical sentence

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

What is Machine Learning?

Automatically uncovering structure in data

	Input	Output
Classification	Vector	Bit (or bits)
Regression	Vector	Real number
Dimensionality reduction	Vector	Shorter vector
Clustering	Vectors	Partition
Sequence labeling	Vectors	Labels
Reinforcement learning	State space + observations	Policy

NLP versus Machine Learning

Machine Translation	Arabic sentence	English sentence
Summarization	Long documents	Short documents
Information Extraction	Document	Database
Question Answering	Corpus + query	Answer
Parsing	Sentence	Syntactic tree
Understanding	Sentence	Logical sentence

Natural Decomposition

Encoding

Ad Hoc Search

Provably Good

Classification	Vector	Bit (or bits)
Regression	Vector	Real number
Dimensionality reduction	Vector	Shorter vector
Clustering	Vectors	Partition
Sequence labeling	Vectors	Labels

Reinforcement learning-inspired	State space	Policy

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

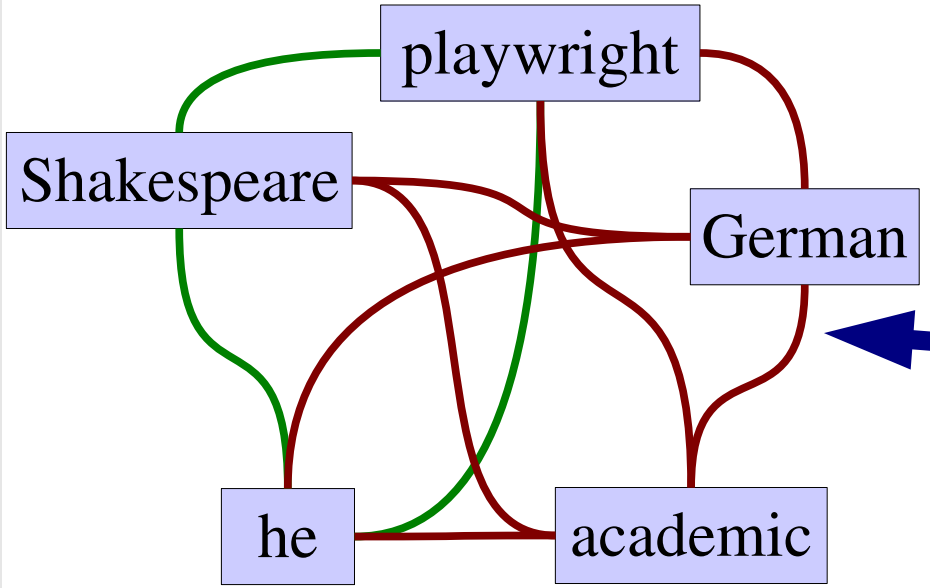
Highly constrained outputs

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.

Shakespeare	scholarship	,	...	a	German	academic	claimed	...
PER	*	*	*		*GPE*	*PER*	*	



Sequence Labeling

Binary Classification + Clustering

Learning in NLP Applications

Model

Features

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:

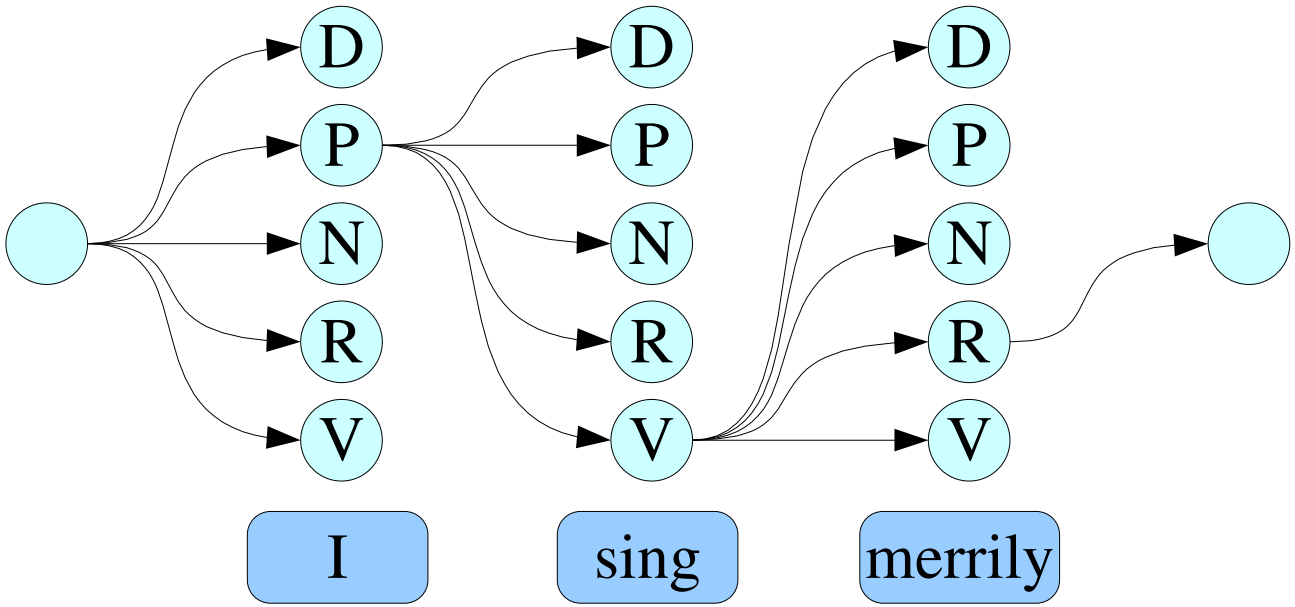
$$\hat{y} = arg\,max_{y \in Y(x)} f(x, y; \theta)$$

Search | Model | Features | Learning

Search (and learning) tractable only for very simple *Y(x)* and *f*

Reduction for Structured Prediction

- Idea: view structured prediction in light of search



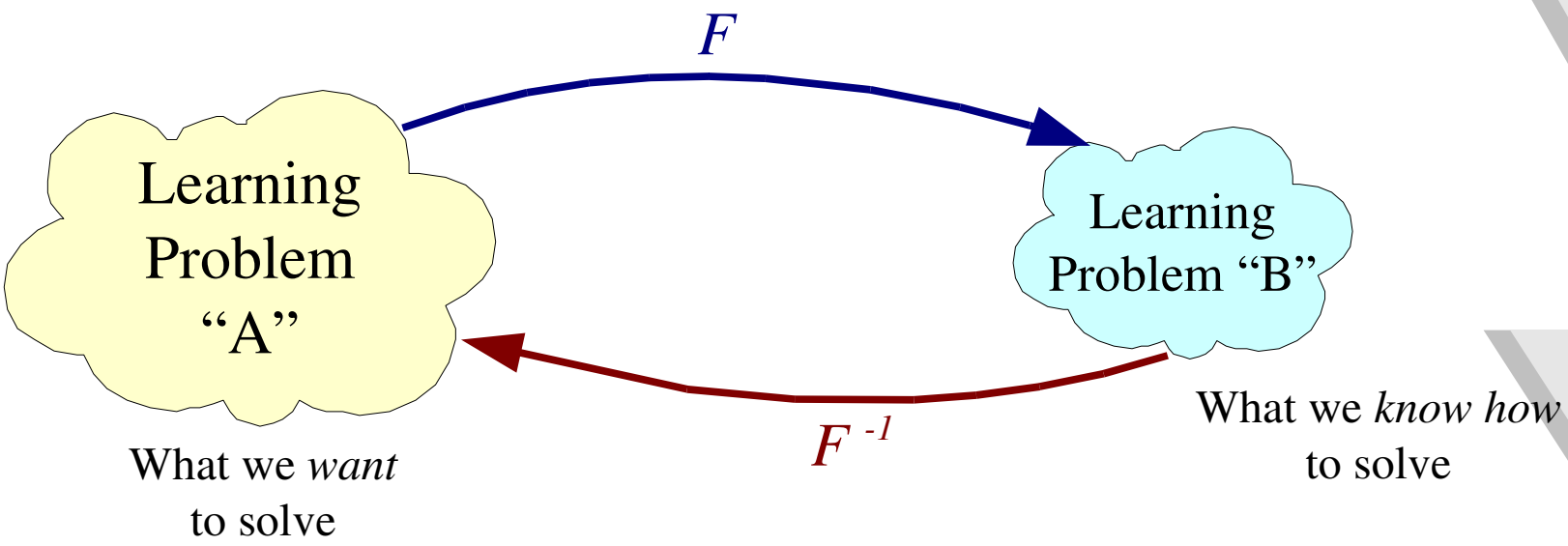
Each step here looks like it could be represented as a weighted multi-class problem.

Can we formalize this idea?

Error-Limiting Reductions

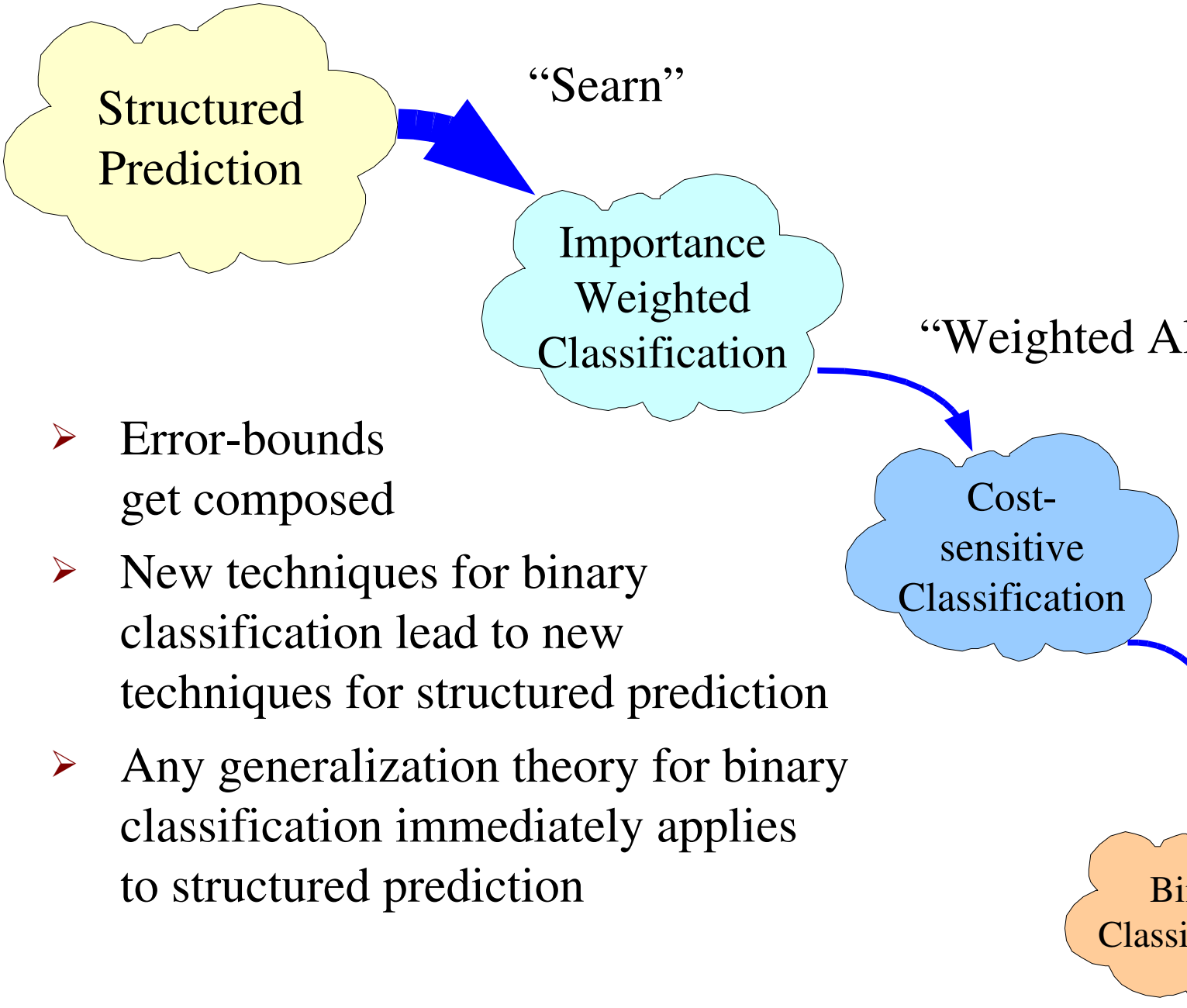
- A formal mapping between learning problems

[Beygelzimer et al., ICML 2005]



- 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

Our Reduction Goal

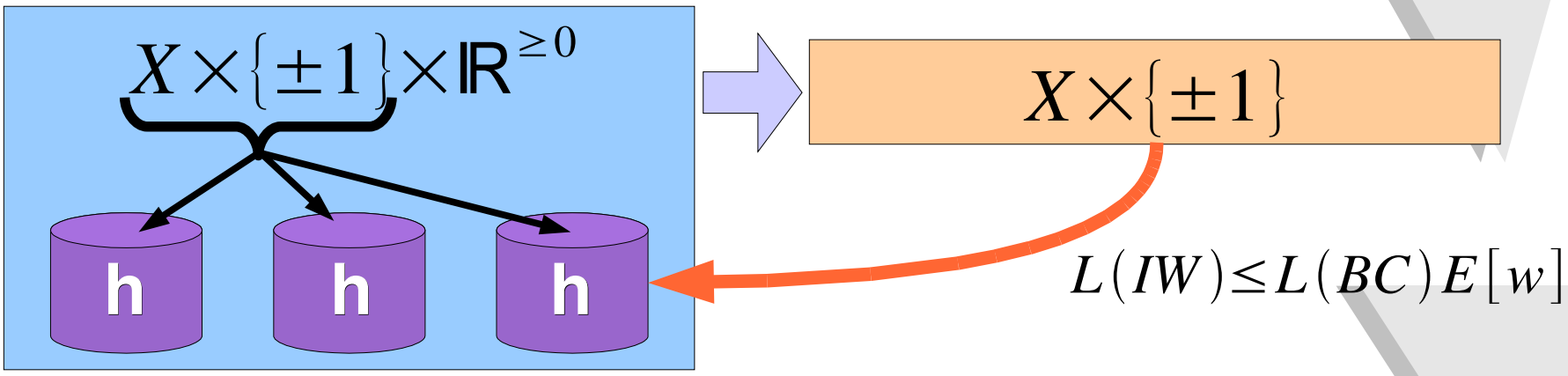


- Error-bounds get composed
- New techniques for binary classification lead to new techniques for structured prediction
- Any generalization theory for binary classification immediately applies to structured prediction

Two Easy Reductions

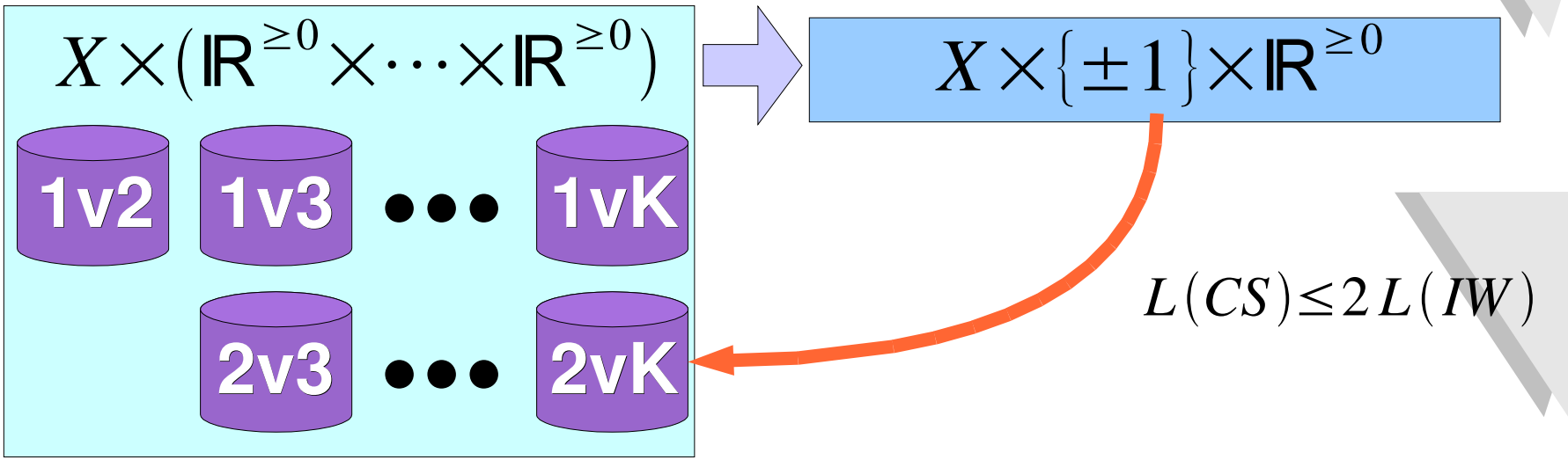
Costing

[Zadronzy, Langford & Abe, ICDM 2003]

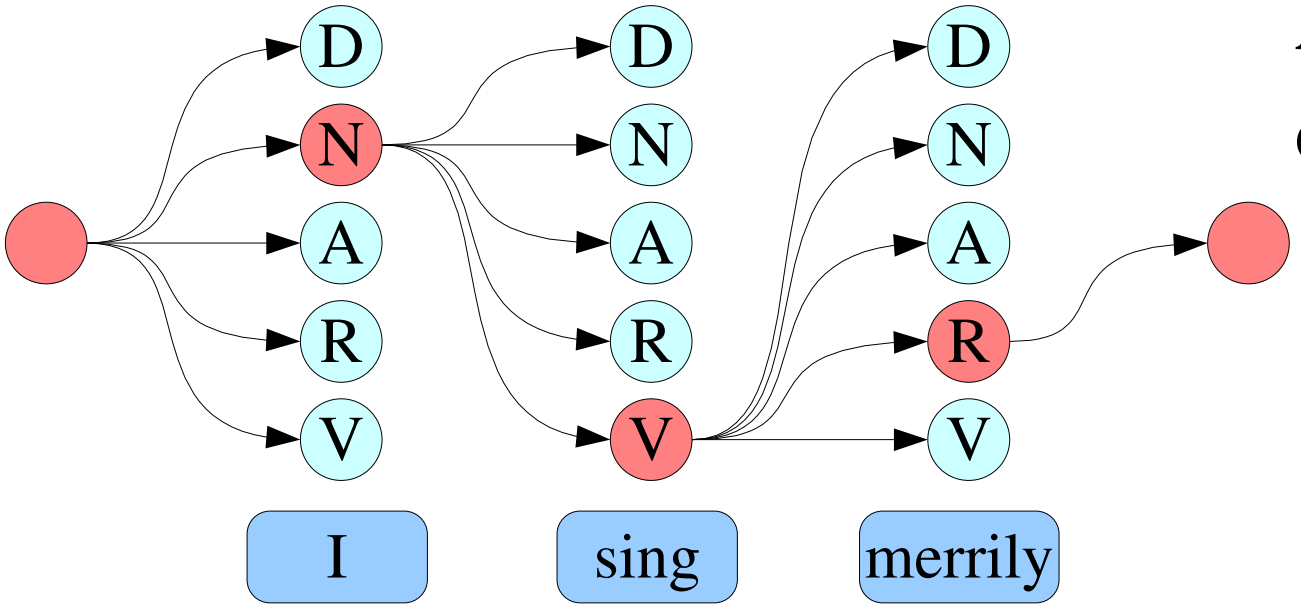


Weighted-All-Pairs

[Beygelzimer, Dani, Hayes, Langford and Zadrozny, ICML 2005]



A First Attempt



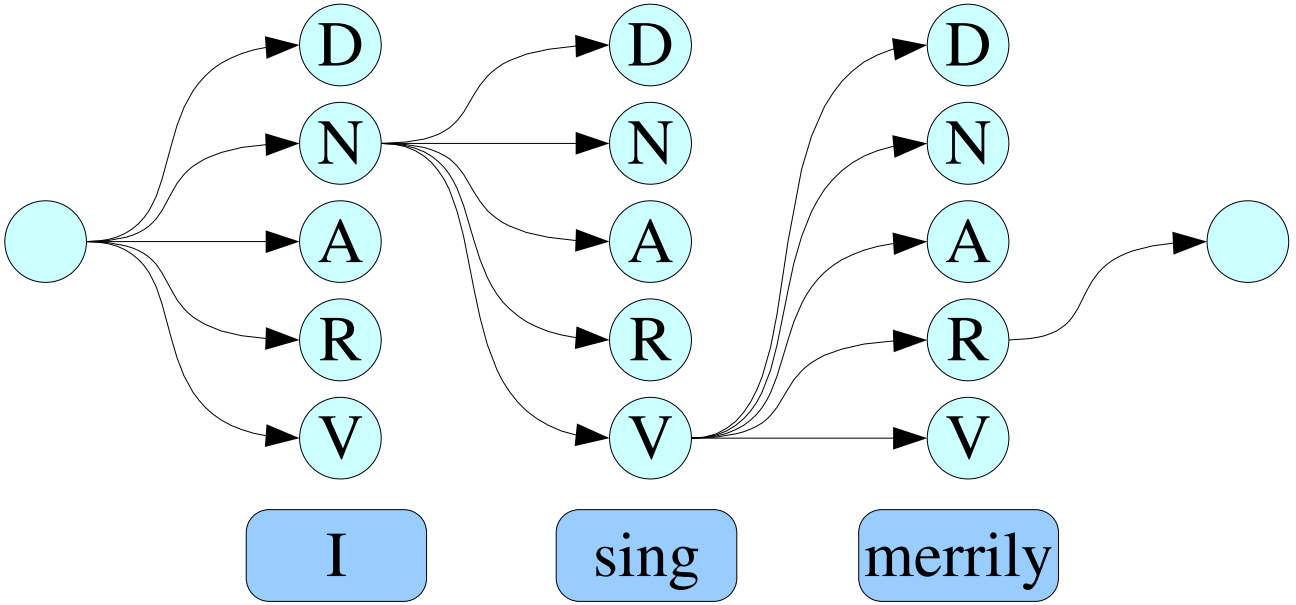
At each correct state:

Train classifier to choose next state

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

$$\frac{T}{2} - \frac{1-(1-2\epsilon)^{T+1}}{4\epsilon} + \frac{1}{2} \approx \frac{T}{2}$$

Reducing Structured Prediction



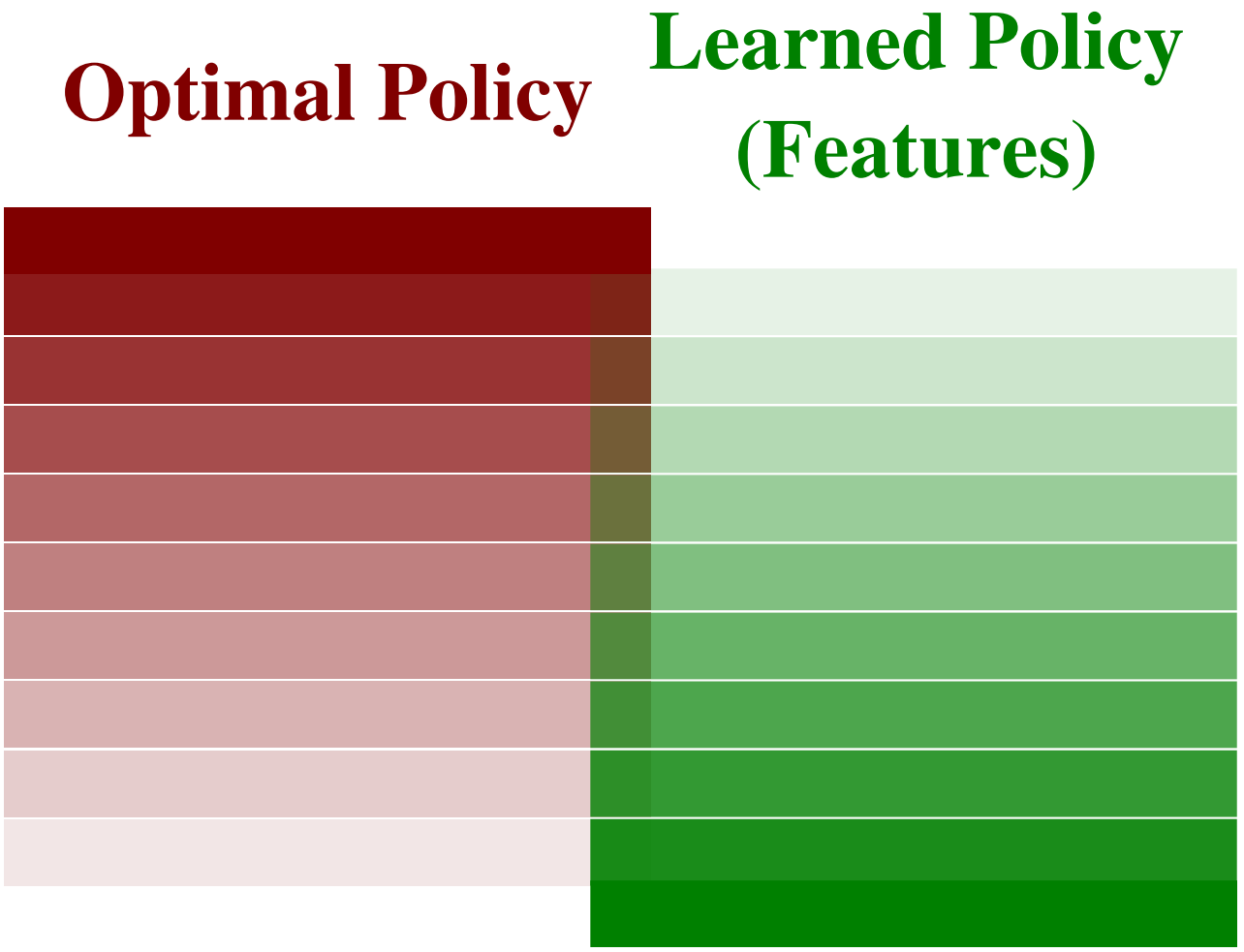
Key Assumption: ***Optimal Policy for training data***

Given: input, true output and state;

Return: best successor state

Weak!

Idea



Iterative Algorithm: Searn

Set current policy = optimal policy

Repeat:

Decode using current policy

... after	a	German	academic	claimed ...	
*	*	**GPE**	**PER**	*	$L = 0$
		PER	*	*	$L = 0.3$
		ORG	**PER**	*	$L = 0.2$
		*	**PER**	*	$L = 0.25$

Generate cl

Learn new multiclass classifier

Interpolate: cur = β*cur + (1-β)*new

Theoretical Analysis

Theorem: For conservative β, after $2T^3 \ln T$ iterations, the loss of the learned policy is bounded as follows:

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

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

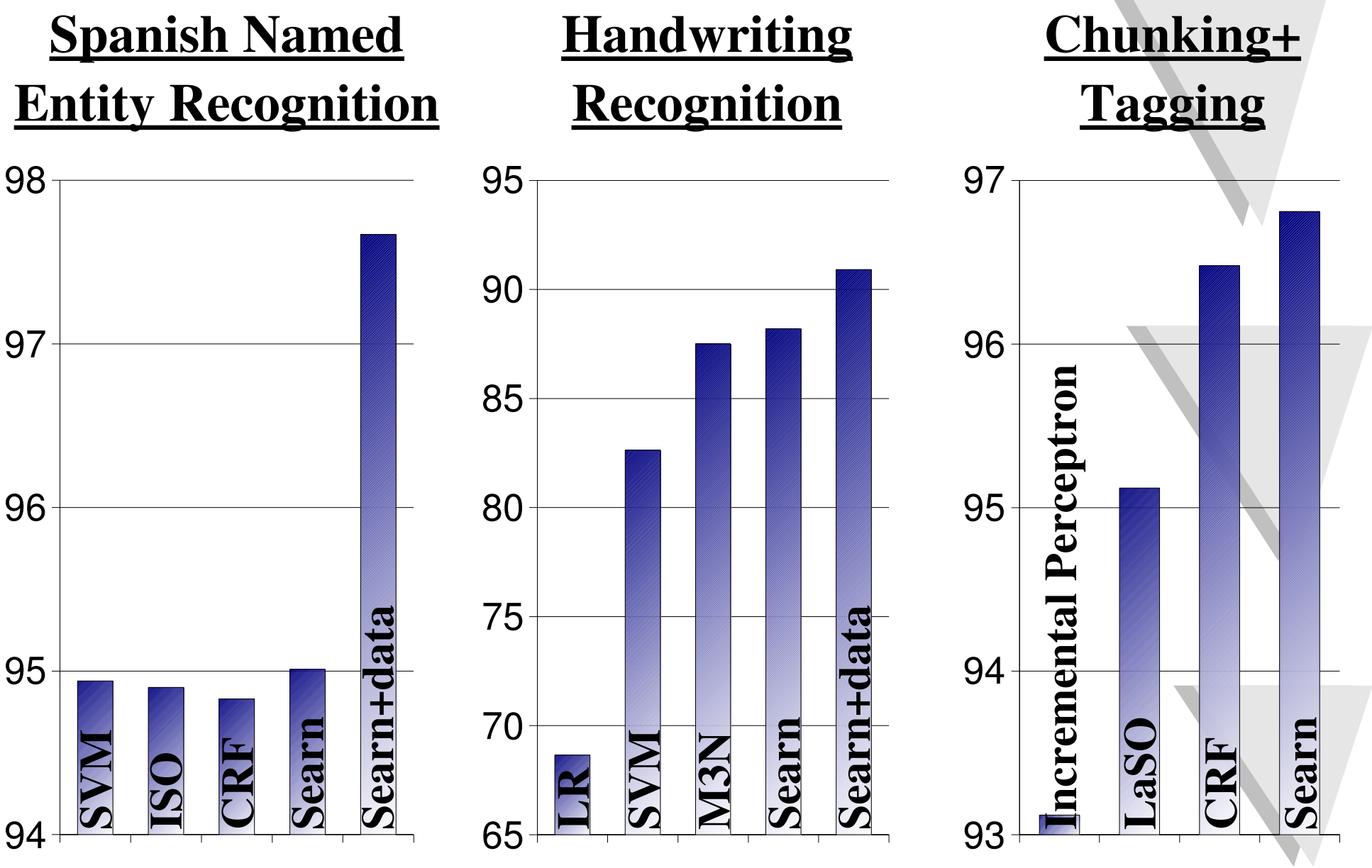
Want to train only where we'll wind up

Searn tells us where we'll wind up

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



Talk Outline

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

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.

... he had died of cancer

... he had died of cancer

... he had died of cancer

... he had died of cancer

... he had died of cancer

... he had died of cancer

... he had died of cancer

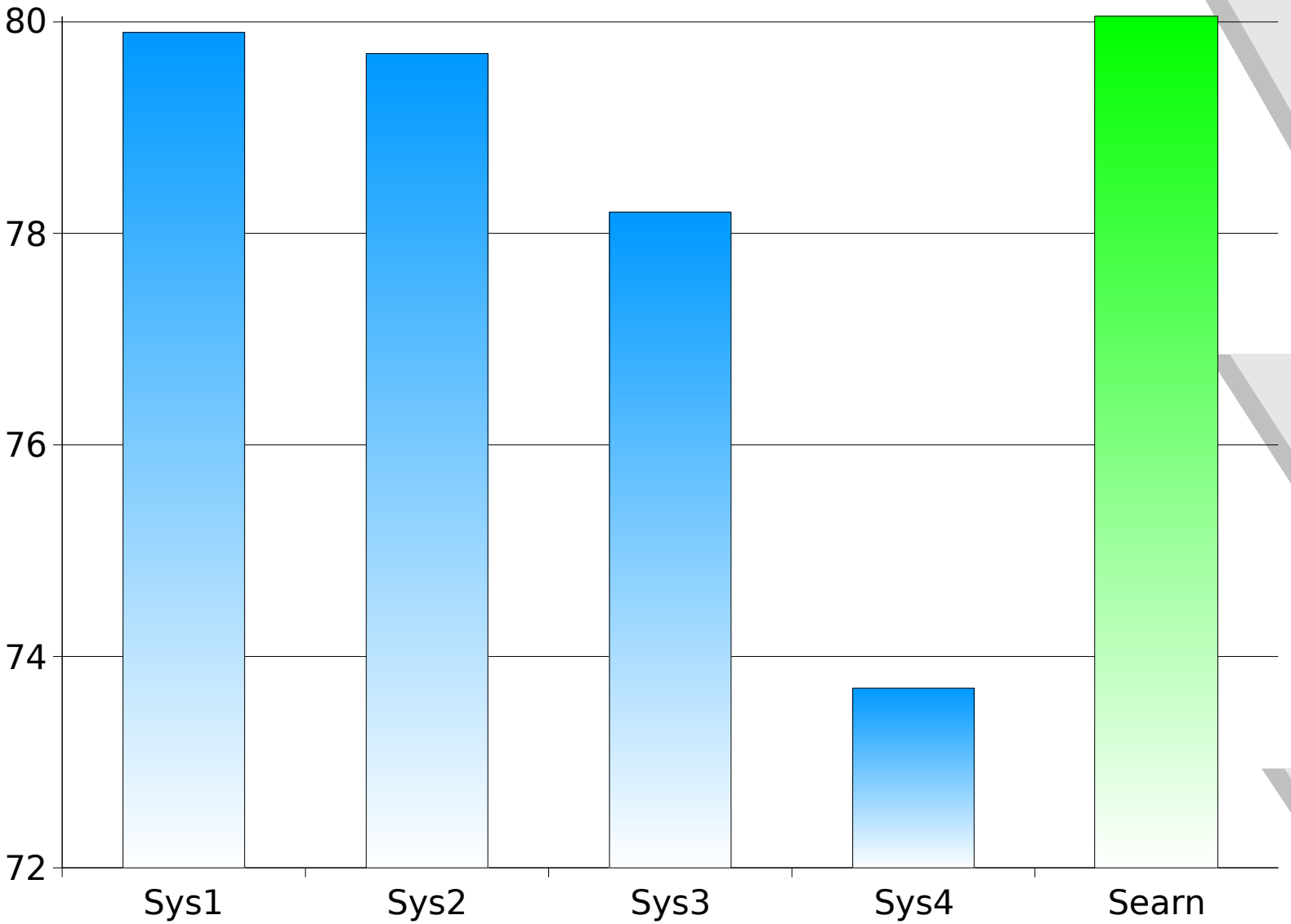
• • •

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.

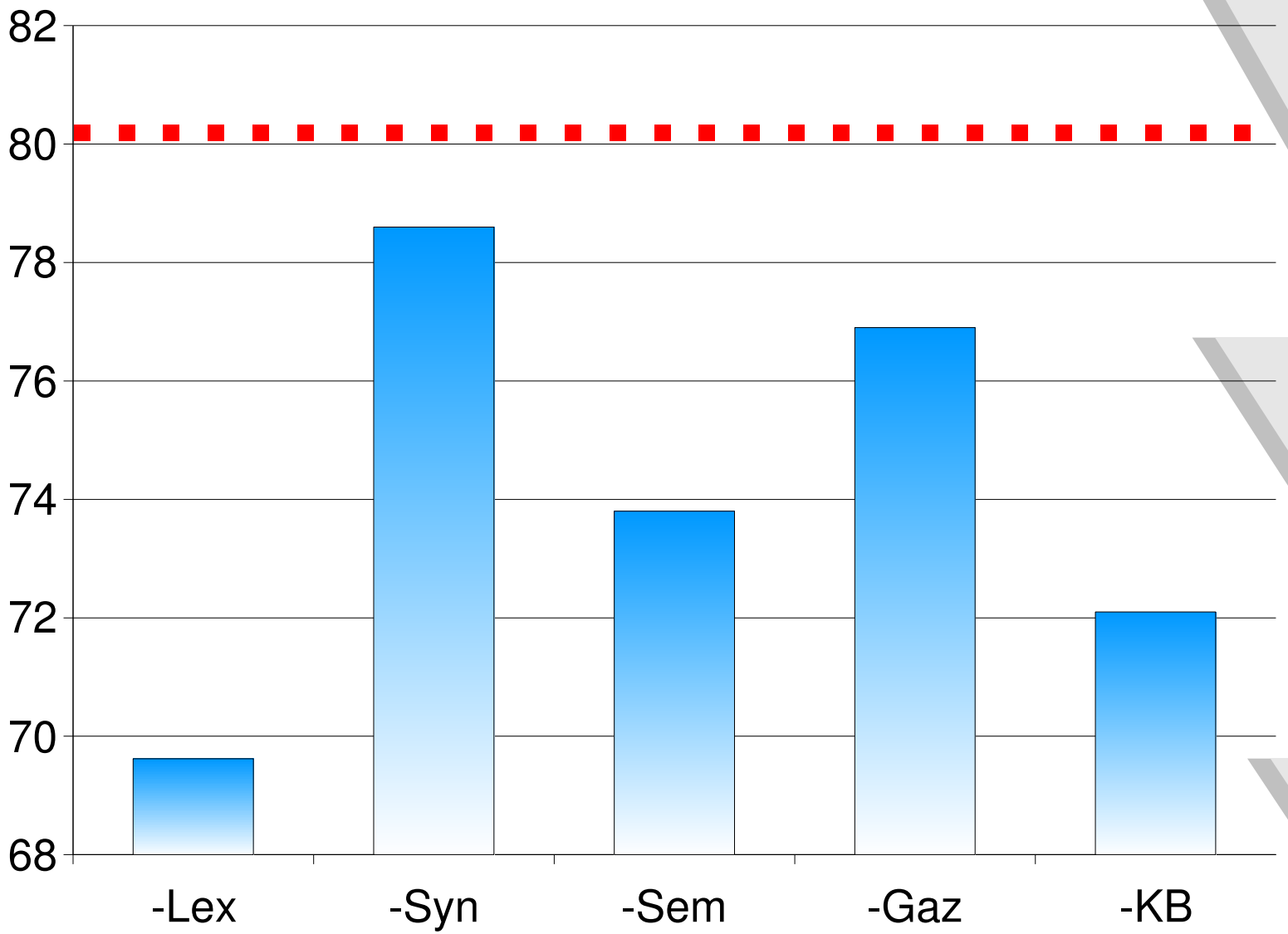
- Lexical
- Syntactic
- Semantic
- Gazetteer
- Knowledge-based

EDT Results (ACE 2004 data)



Previous state of the art
with extra proprietary data

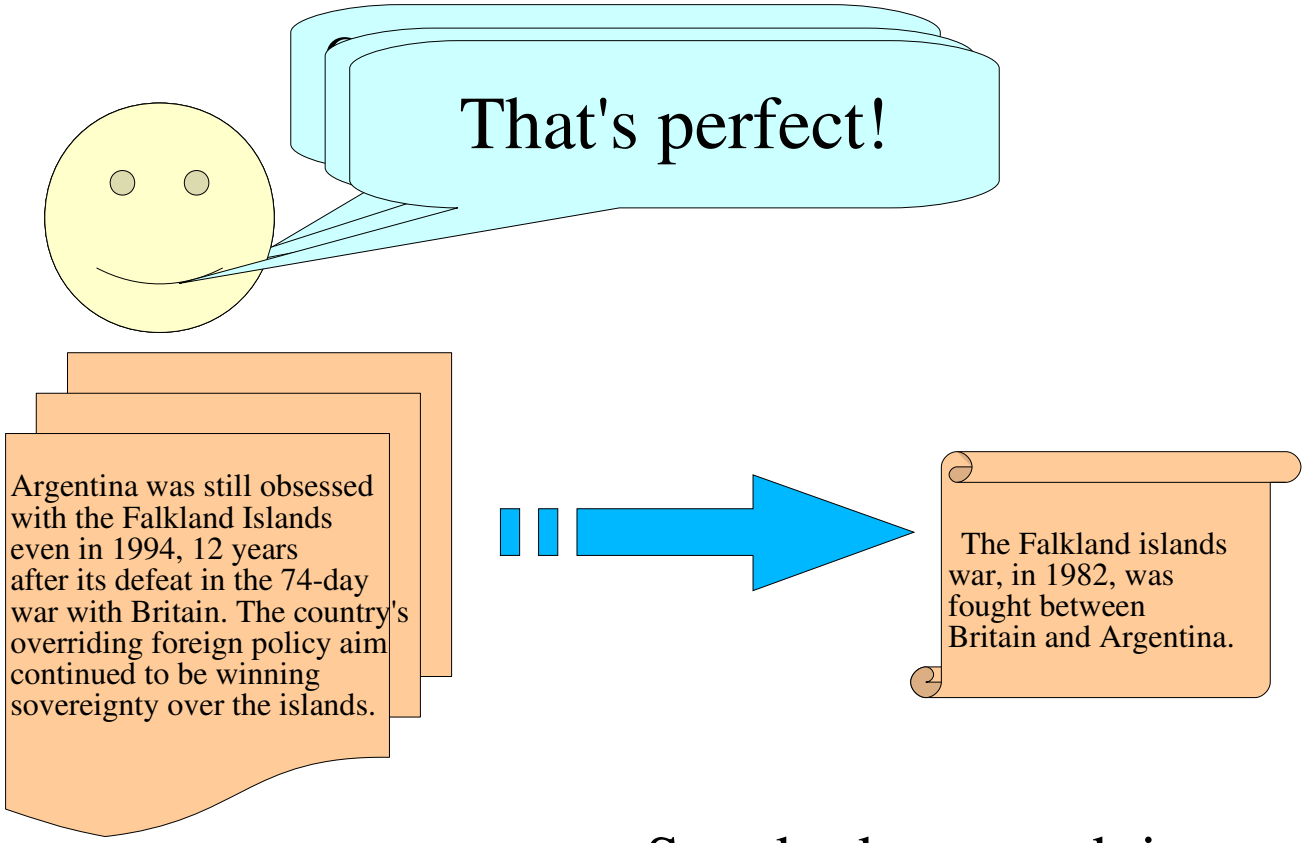
Feature Contributions



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



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

Structure of Search

To make search more tractable, we run an initial round of sentence extraction @ 5x length

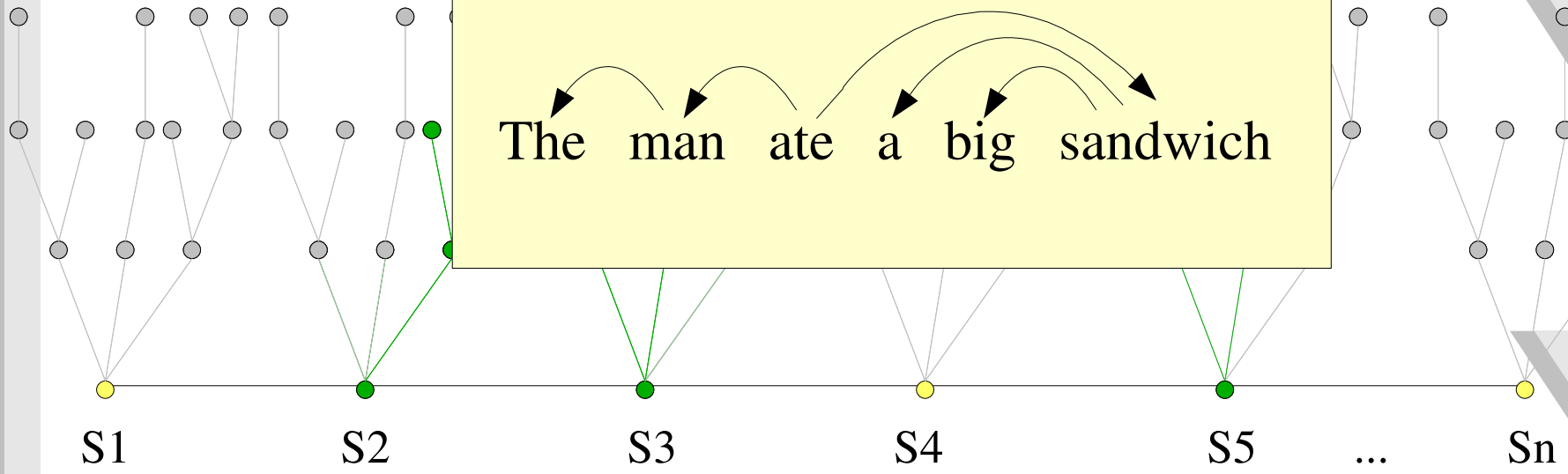
Argentina was still obsessed with the Falkland Islands even in 1994, 12 years after its defeat in the 74-day war with Britain. The country's overriding foreign policy aim continued to be winning sovereignty over the islands.

ummary

The a big

ndwich

Optimal Policy not analytically available; Approximated with beam search.



= frontier node = summary node

Example Output (40 word limit)

Sentence Extraction + Compression:

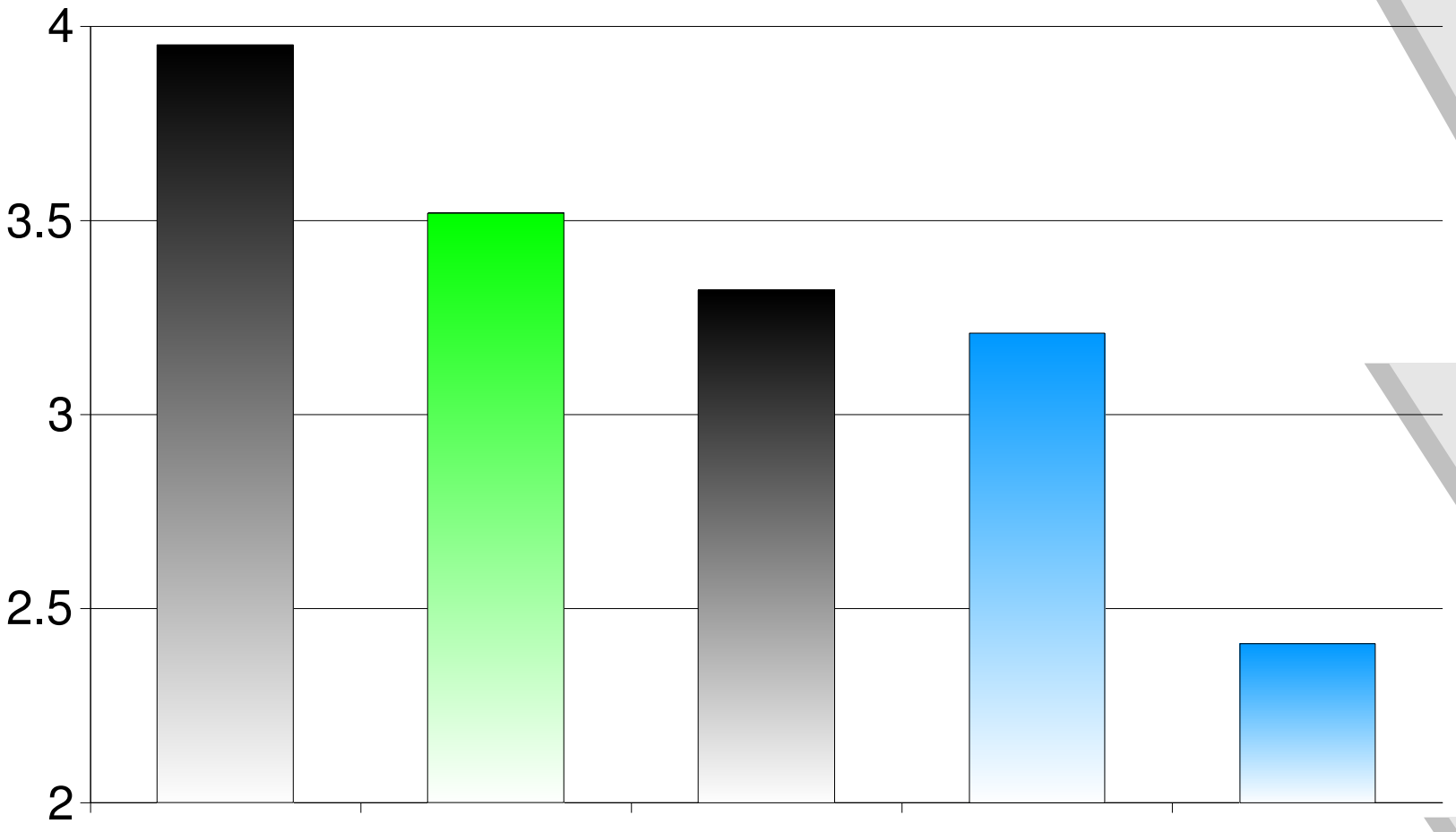
+13 Argentina and Britain announced an agreement, nearly eight years after they fought a 74-day war a populated archipelago off Argentina's coast. Argentina gets out the red carpet, official royal visitor since the end of the Falklands war in 1982.

Vine Growth:

+24 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	Diplomatic ties restored	3	Falkland war was in 1982
5	Major cabinet member visits	3	Cavallo visited UK
5	Exchanges were in 1992	2	War was 74-days long
3	War between Britain and Argentina		

Results



[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.

What Can Searn Not Do? (Yet)

Solve structured prediction problems...

~~...with only weak feedback.~~

~~...with hidden variables.~~

...with enormous branching factors.

...without pre-defined search spaces.

...with "combinatorial" outputs.

...with only an optimal path.

Thanks!

Questions?

BACKUP SLIDES

Optimal Policies

Weak Feedback

?∪

Optimal Path

?∪

Approximately Optimal Policy

∪

Optimal Policy

} Most NLP Problems

Searn
[Daumé III, Langford, Marcu, submitted to NIPS 2006]

∪

Optimally Completable

∪

Incremental Perceptron
[Collins & Roark, EMNLP 2004]
LaSO
[Daumé III, Marcu, ICML 2005]

Loss-augmented Search

∪

Max-margin Markov Networks
[Taskar, Guestrin, Koller, NIPS 2003]

Normalized Search

Conditional Random Fields
[Lafferty, McCallum, Pereira, ICML 2001]