Any Domain Parsing

Automatic Domain Adaptation for Natural Language Parsing

David McClosky

Department of Computer Science
Brown University

September 18th, 2009

Committee: Eugene Charniak, Mark Johnson, and Dan Klein

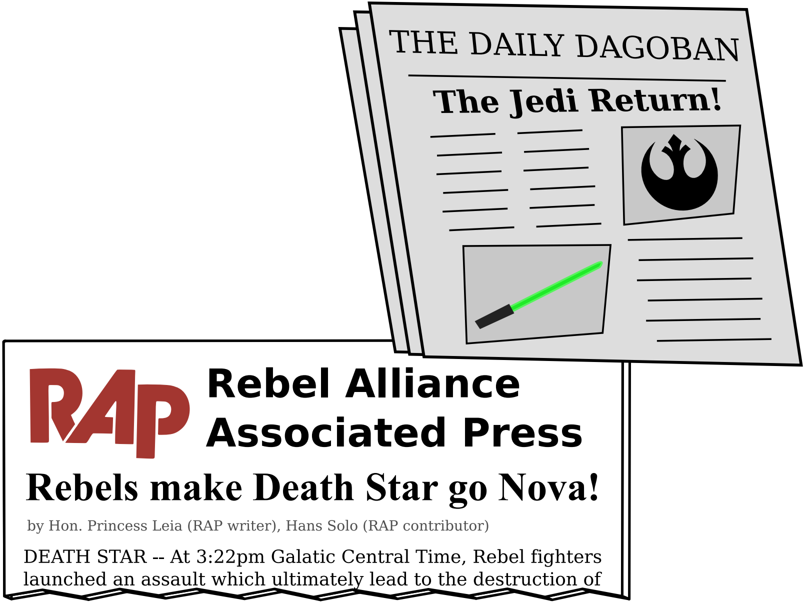
Robots will need to understand language

Robots will need to understand language

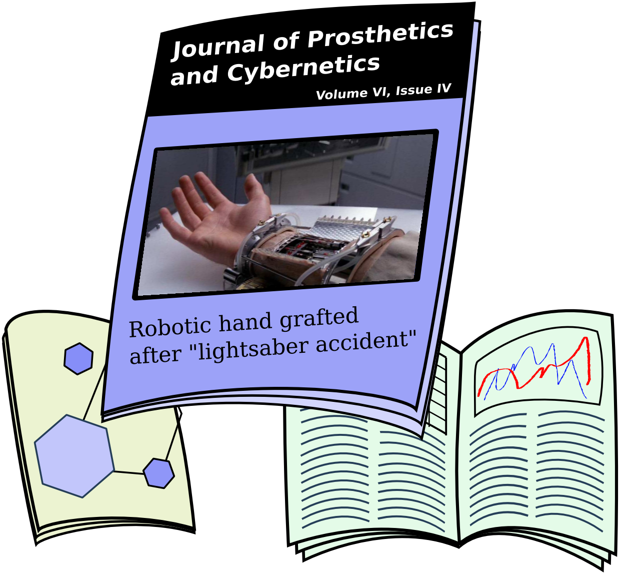
[Lucas et al., 1977, Lucas et al., 1980, Lucas et al. 1983]

Keeping up to date with Twitter

Reading the news

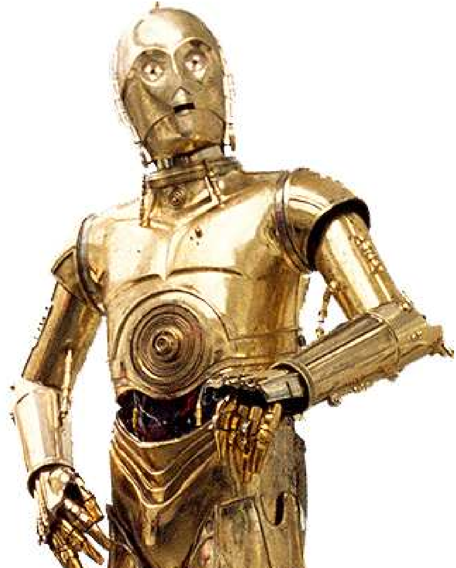


Studying the latest medical journals

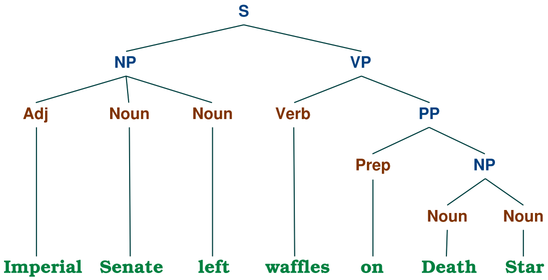


Casual reading

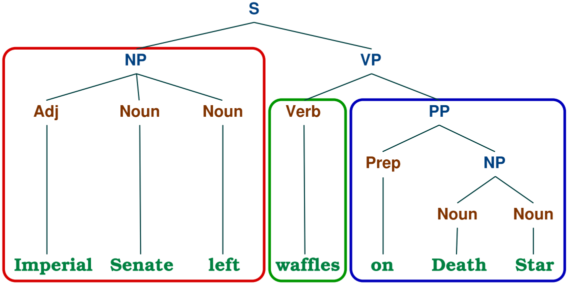
Imperial Senate left waffles on Death Star.



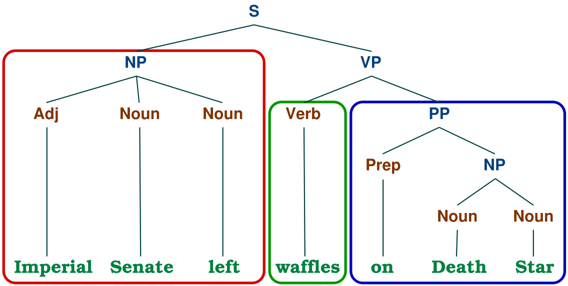
Semantics depend on syntax

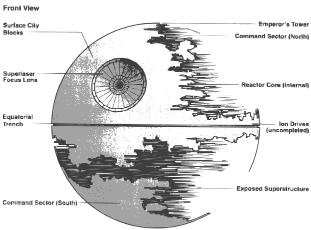


Semantics depend on syntax

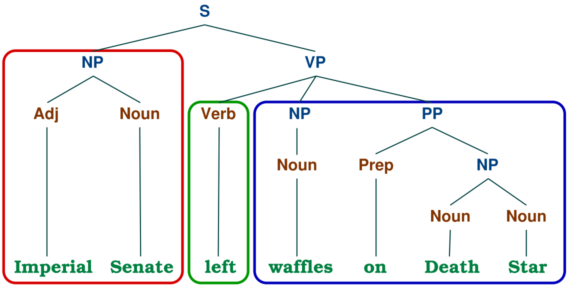


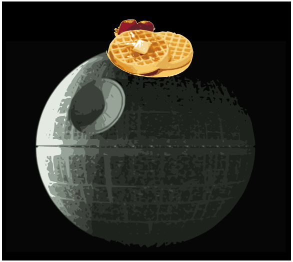
Semantics depend on syntax





Semantics depend on syntax





Applications of Parsing

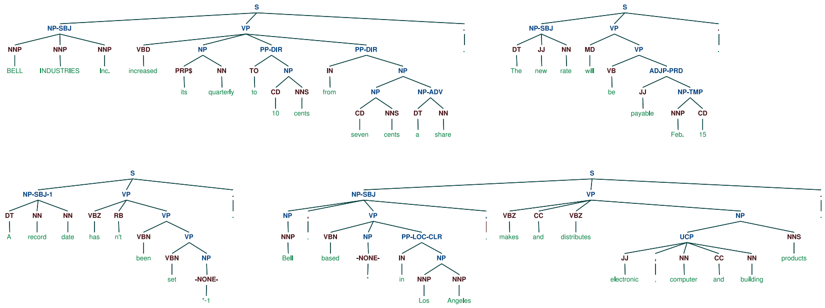
- Parsing is often part of larger NLP pipelines.
- Some examples:
  - Machine translation [Charniak *et al.*, 2003]
  - Bioinformatics [Miyao *et al.*, 2008]
  - Forensics (author identification) [Luyckx and Daelemans, 2008]
  - Discourse analysis [Barzilay and Lapata, 2008]
  - Summarization [Turner and Charniak, 2005]
  - Language modeling [Roark, 2001], [Charniak, 2001]
  - Speech repairs [Johnson and Charniak, 2004]
  - Coreference [Luo and Zitouni, 2005], [Charniak and Elsner, 2009]
  - etc.

Data-driven Parsing

- Many current approaches to parsing are data-driven.

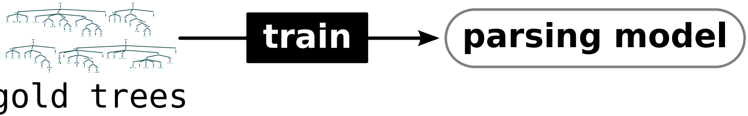
Data-driven Parsing

- Many current approaches to parsing are data-driven.
- Data consists of human-annotated corpora with labeled examples of correct parse structures (“gold trees”).



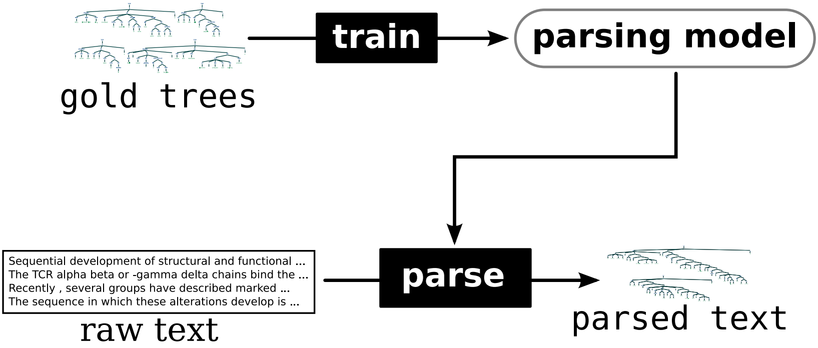
Data-driven Parsing

- Parsers are trained on these corpora to produce models.



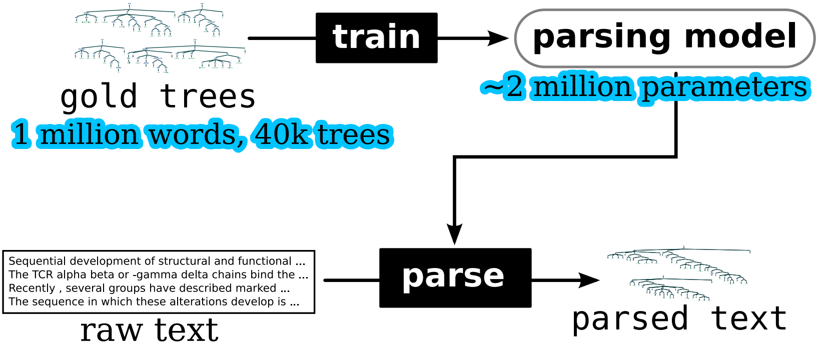
Data-driven Parsing

- The parsing model is used to parse unlabeled text.



Data-driven Parsing

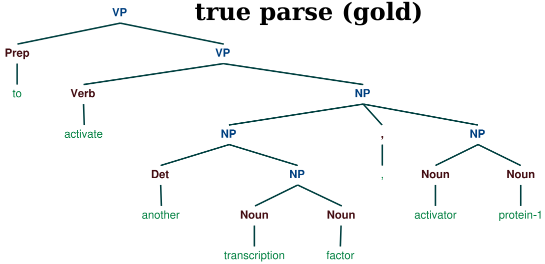
- Many model parameters may hurt portability/generality



What’s in a domain?

Masters of their domain?

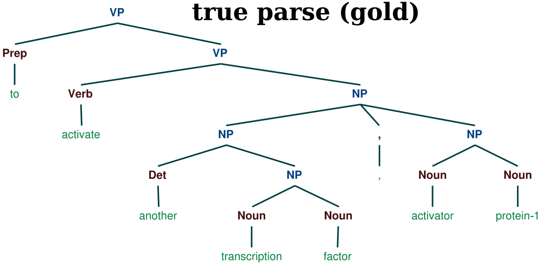
true parse (gold)

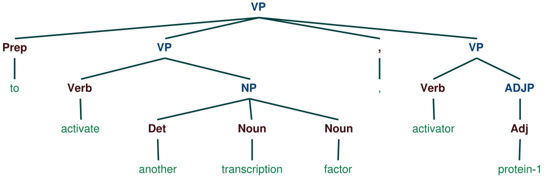


Masters of their domain?

PubMed

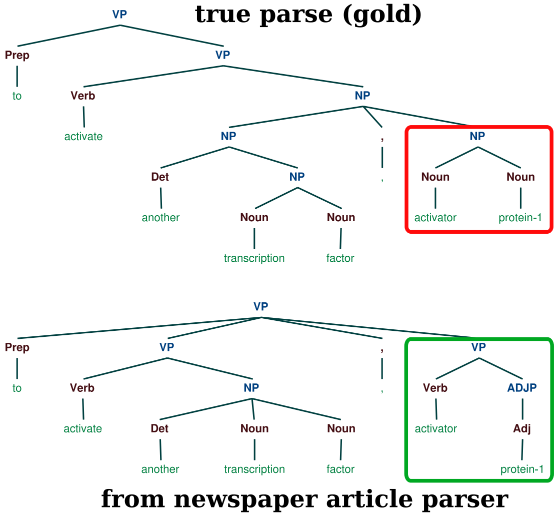
true parse (gold)





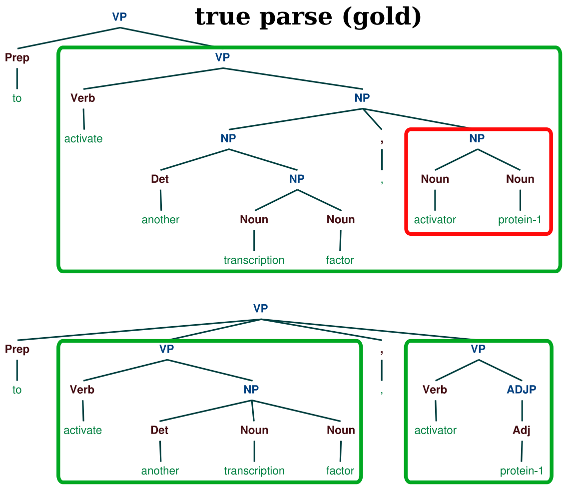
from newspaper article parser

Masters of their domain?



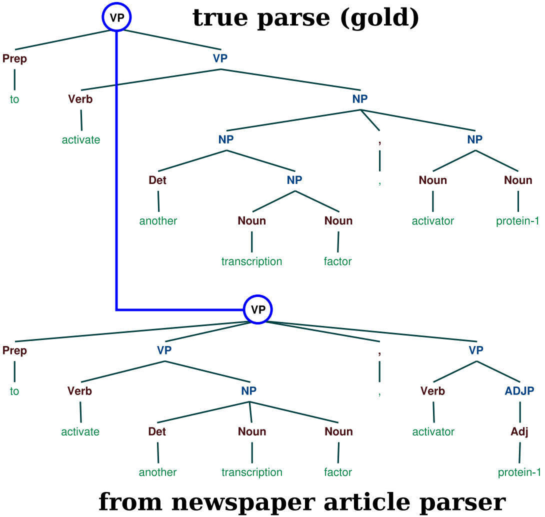
Masters of their domain?

PubMed

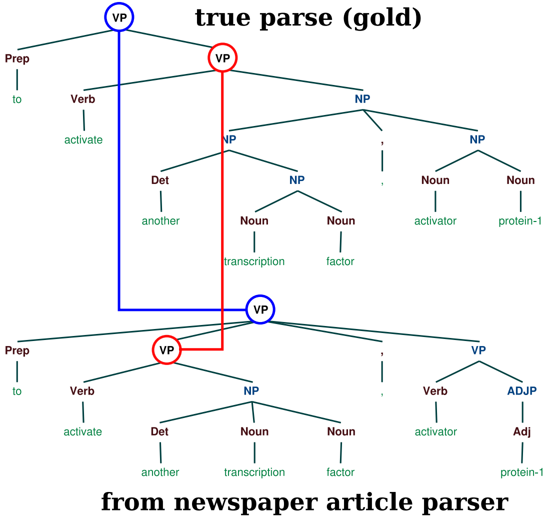


from newspaper article parser

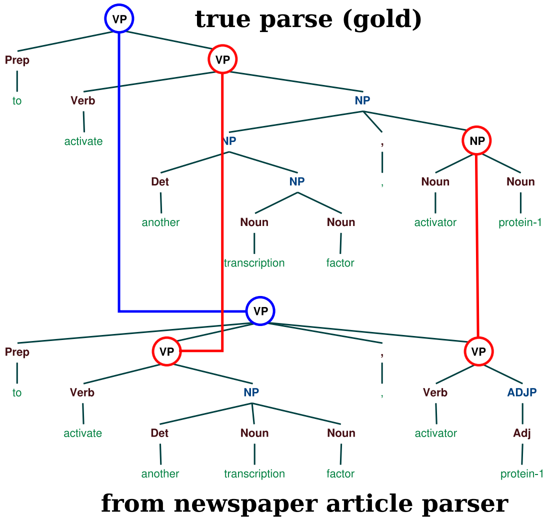
Evaluating parse trees



Evaluating parse trees

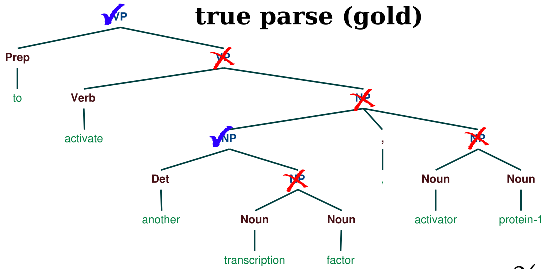


Evaluating parse trees



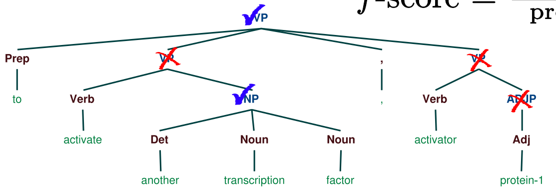
Evaluating parse trees

true parse (gold)



$$\text{recall} = \frac{2}{6}$$

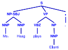
$$f\text{-score} = \frac{2(\text{precision})(\text{recall})}{\text{precision}+\text{recall}} = \frac{4}{11}$$



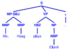
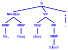
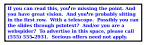
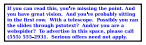
$$\text{precision} = \frac{2}{5}$$

from newspaper article parser

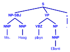
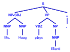
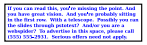
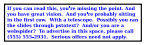
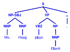
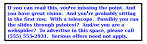
Parsing training scenarios

Train	Test	Description
	If you can read this, you're missing the point. And you have great vision. And you're probably sitting in the first row. With a telescope. Possibly you ran the slides through pdftotext? And/or you are a webspider? To advertise in this space, please call (555) 555-2931. Serious offers need not apply.	supervised parsing

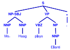
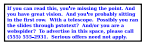
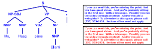
Parsing training scenarios

Train	Test	Description
		supervised parsing
 		semi-supervised parsing

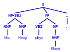
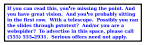
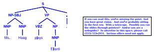
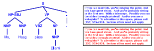
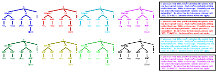
Parsing training scenarios

Train	Test	Description
		supervised parsing
 		semi-supervised parsing
 		parser portability

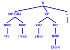
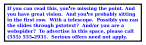
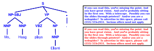
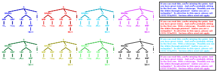
Parsing training scenarios

Train	Test	Description
		supervised parsing
		semi-supervised parsing
		parser portability
		parser adaptation

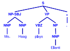
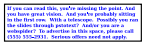
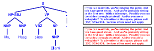
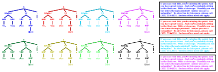
Parsing training scenarios

Train	Test	Description
		supervised parsing
		semi-supervised parsing
		parser portability
		parser adaptation
		multiple source parser adaptation

Parsing training scenarios

Train	Test	Description
		supervised parsing
		semi-supervised parsing
		parser portability
		parser adaptation
		multiple source parser adaptation

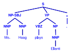
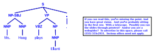
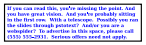
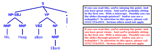
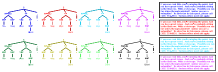
Parsing training scenarios

Train	Test	Description
		supervised parsing
		semi-supervised parsing
		parser portability
		parser adaptation
		multiple source parser adaptation

Thesis statement

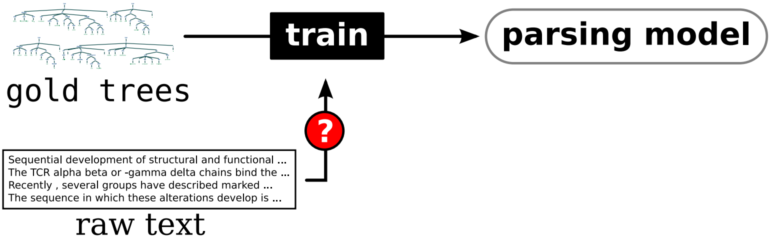
Self-training is an effective semi-supervised learning technique for parsing, capable of improving both in-domain and cross-domain parsing scenarios.

Parsing training scenarios

Train	Test	Description
		supervised parsing
		semi-supervised parsing
		parser portability
		parser adaptation
		multiple source parser adaptation

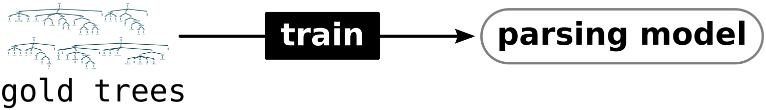
Incorporating unlabeled data

- How can we leverage unlabeled data in our models?



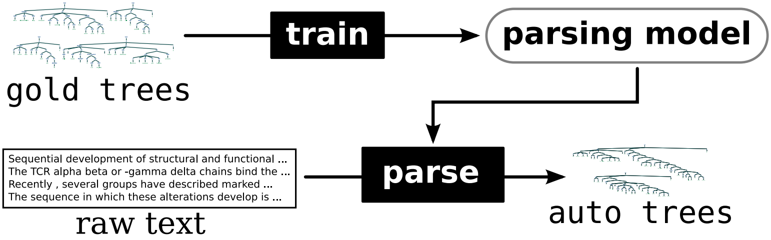
Incorporating unlabeled data: Self-training

1. Train a model from the labeled data.



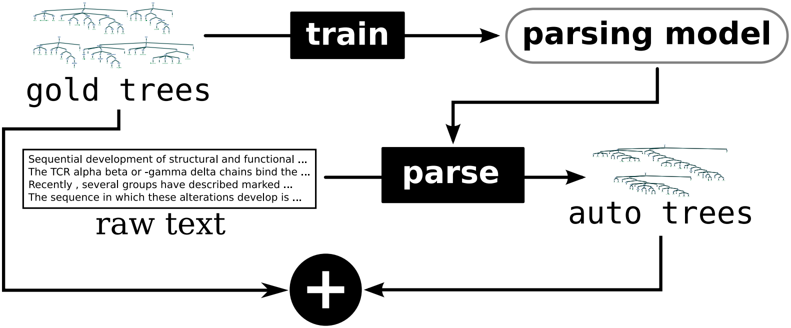
Incorporating unlabeled data: Self-training

2. Parse the unlabeled text.



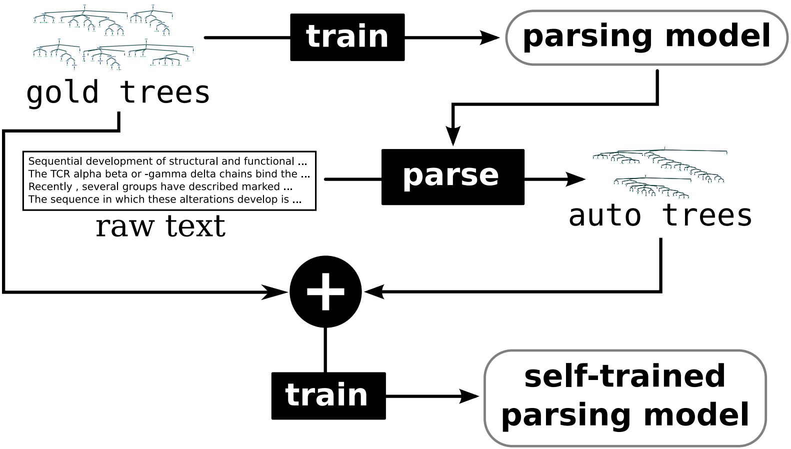
Incorporating unlabeled data: Self-training

3. Combine gold trees with automatically parsed trees.



Incorporating unlabeled data: Self-training

4. Train a new model from the combination.



A Brief History of Self-training

- [Charniak, 1997]
- [Steedman *et al.*, 2003]
- [Clark and Curran, 2003] (part of speech tagging)
- [Roark and Bacchiani, 2003]

A Brief History of Self-training

- [Charniak, 1997]
- [Steedman *et al.*, 2003]
- [Clark and Curran, 2003] (part of speech tagging)
- [Roark and Bacchiani, 2003]

  $\rightarrow$ no improvements over state-of-the-art from self-training

A Brief History of Self-training

- [Charniak, 1997]
- [Steedman *et al.*, 2003]
- [Clark and Curran, 2003] (part of speech tagging)
- [Roark and Bacchiani, 2003]

  → no improvements over state-of-the-art from self-training

- [McClosky, Charniak, and Johnson, NAACL 2006]

A Brief History of Self-training

- [Charniak, 1997]
- [Steedman *et al.*, 2003]
- [Clark and Curran, 2003] (part of speech tagging)
- [Roark and Bacchiani, 2003]

  → no improvements over state-of-the-art from self-training

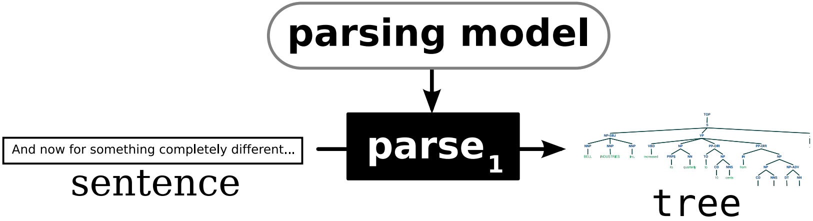
- [McClosky, Charniak, and Johnson, NAACL 2006]

  → reranking parser

The Parsing Model

[Charniak and Johnson, 2005]

- Lexicalized PCFG parser gives most probable parse



“first stage parser”

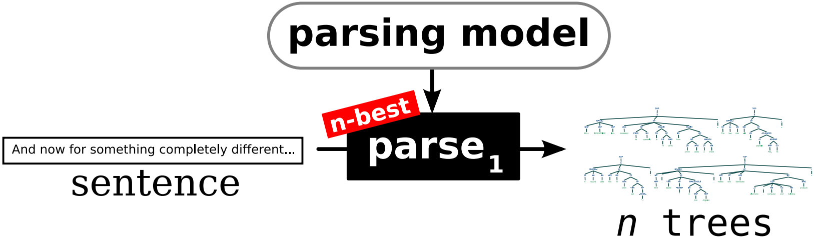
or

“parser”

The Parsing Model

[Charniak and Johnson, 2005]

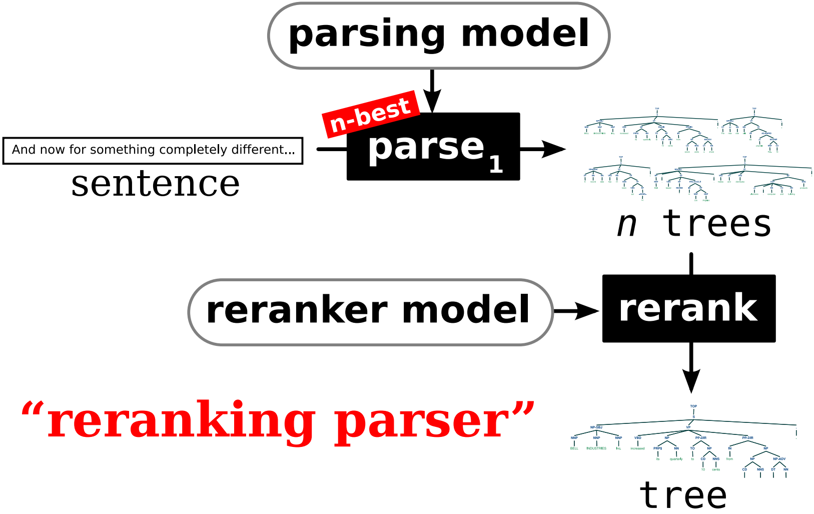
- Use *n* most probable parses instead just top parse



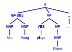
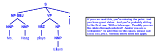
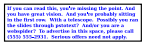
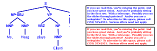
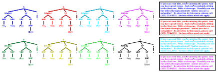
The Parsing Model

[Charniak and Johnson, 2005]

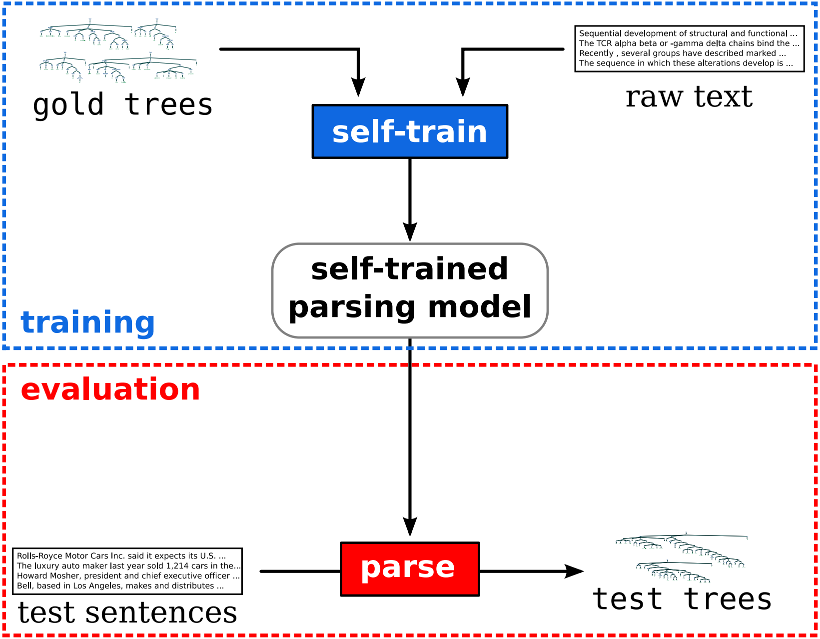
- Discriminative reranker picks “best” parse from list



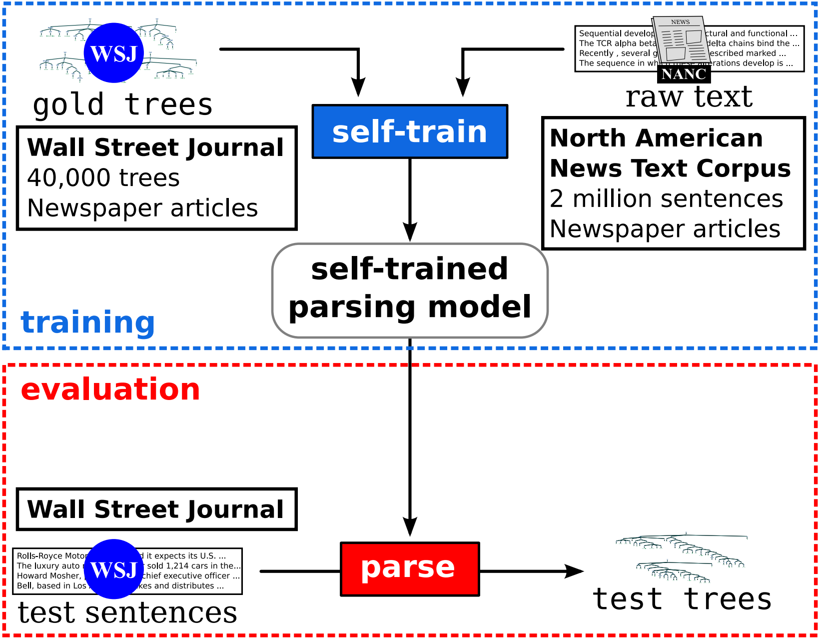
Parsing training scenarios

Train	Test	Description
		supervised parsing
		semi-supervised parsing
		parser portability
		parser adaptation
		multiple source parser adaptation

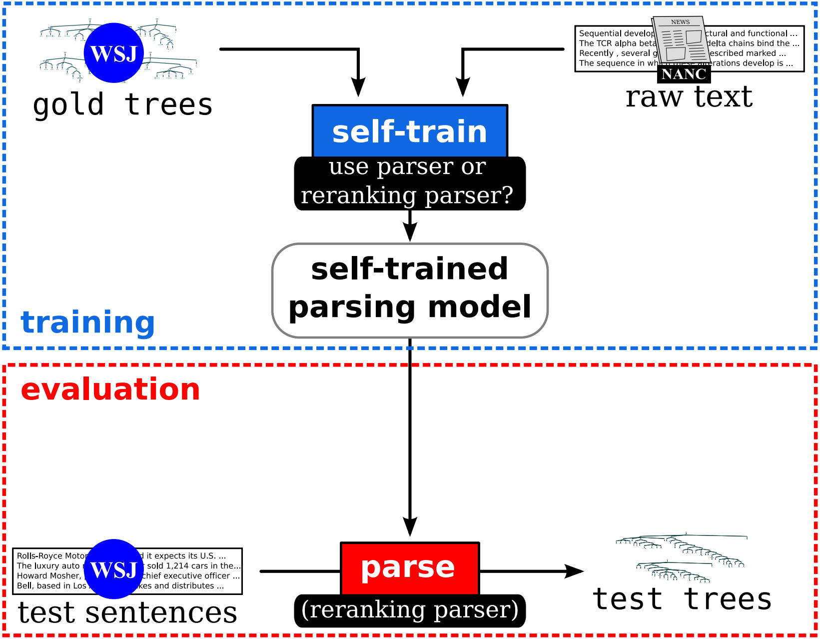
Self-training for parsing



Self-training for parsing



Self-training for parsing



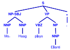
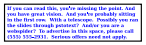
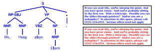
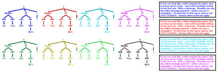
Self-training for parsing is effective

[McClosky, Charniak, and Johnson, NAACL 2006]

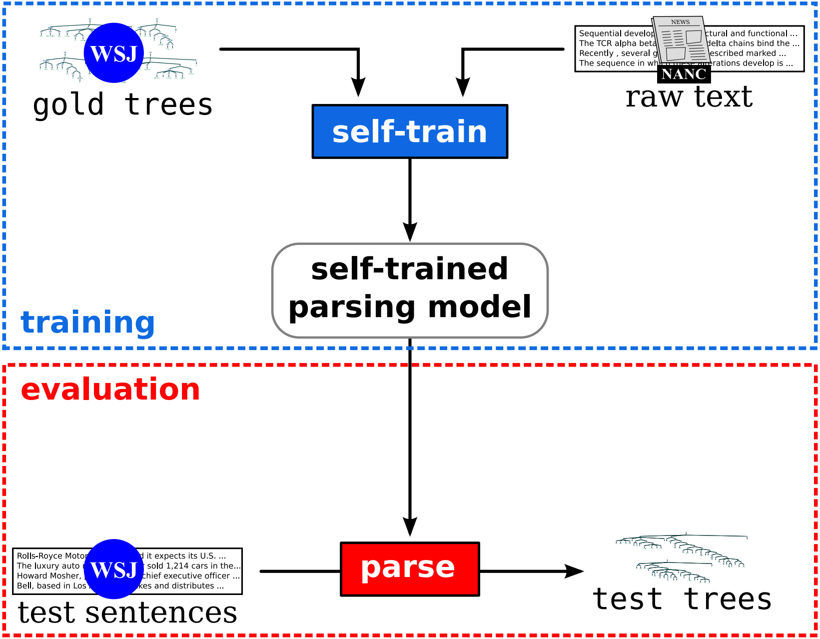
Model	*f*-score
Baseline (WSJ)	91.3
Self-trained (WSJ + NANC)	92.1

f-scores on WSJ evaluation section

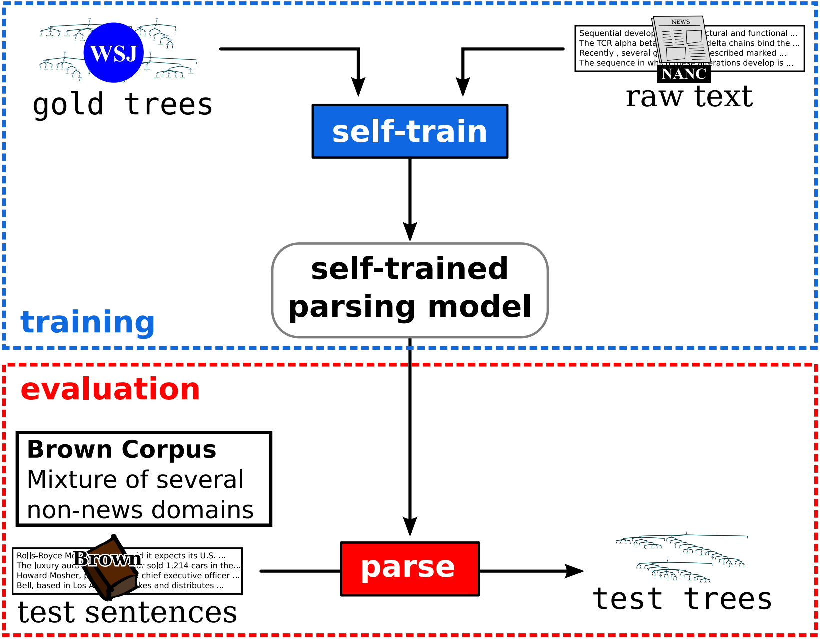
Parsing training scenarios

Train	Test	Description
		supervised parsing
		semi-supervised parsing
		parser portability
		parser adaptation
		multiple source parser adaptation

Parser portability experiments



Parser portability experiments



Self-trained WSJ model portability

Train	Test	*f*-score
WSJ	WSJ	91.3
WSJ	BROWN	85.2

f-score on WSJ and BROWN evaluation sections

Self-trained WSJ model portability

Train	Test	*f*-score
WSJ	WSJ	91.3
WSJ	BROWN	85.2
WSJ + NANC	BROWN	87.8

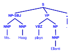
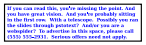
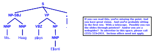
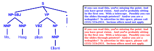
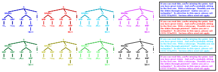
f-score on WSJ and BROWN evaluation sections

Self-trained WSJ model portability

Train	Test	*f*-score
WSJ	WSJ	91.3
WSJ	BROWN	85.2
WSJ + NANC	BROWN	87.8
BROWN	BROWN	88.4

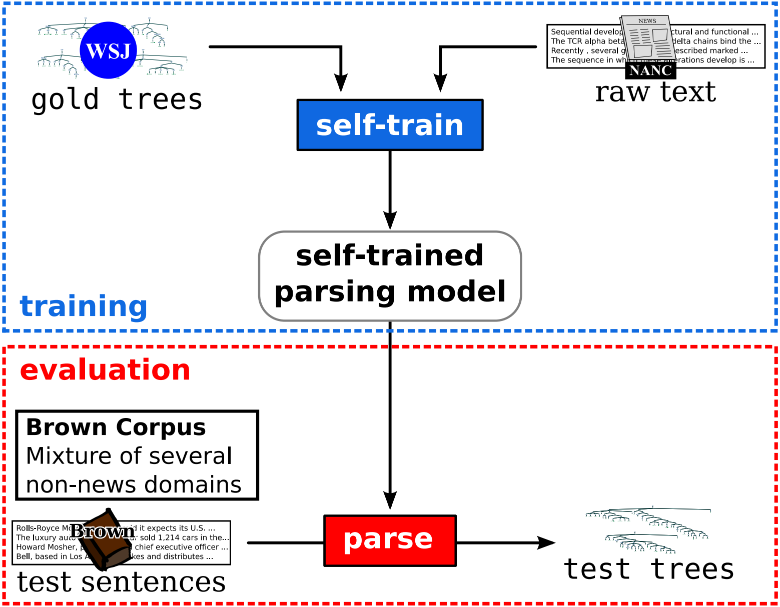
f-score on WSJ and BROWN evaluation sections

Parsing training scenarios

Train	Test	Description
		supervised parsing
		semi-supervised parsing
		parser portability
		parser adaptation
		multiple source parser adaptation

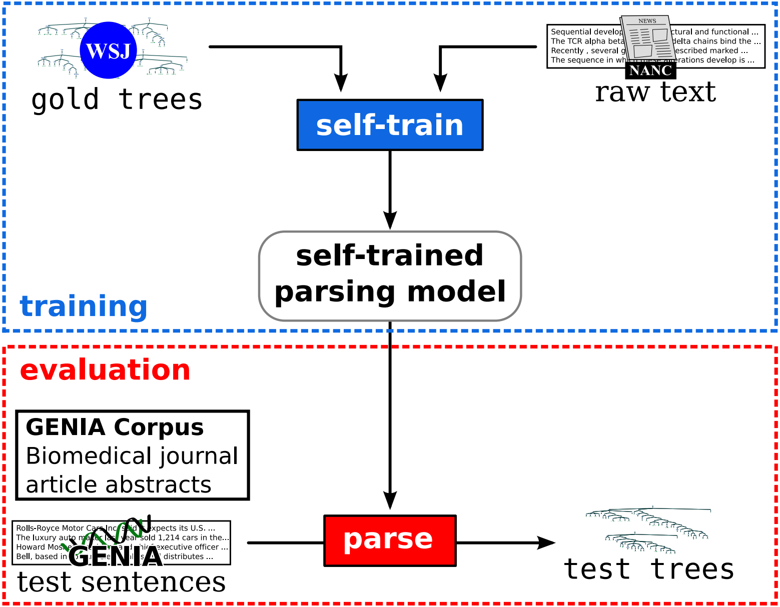
Parser adaptation experiments

More distant domains...



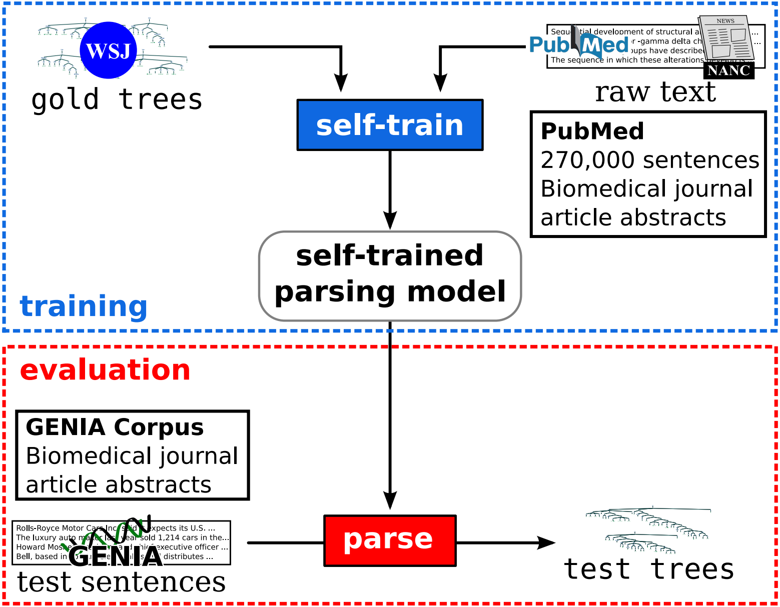
Parser adaptation experiments

More distant domains...



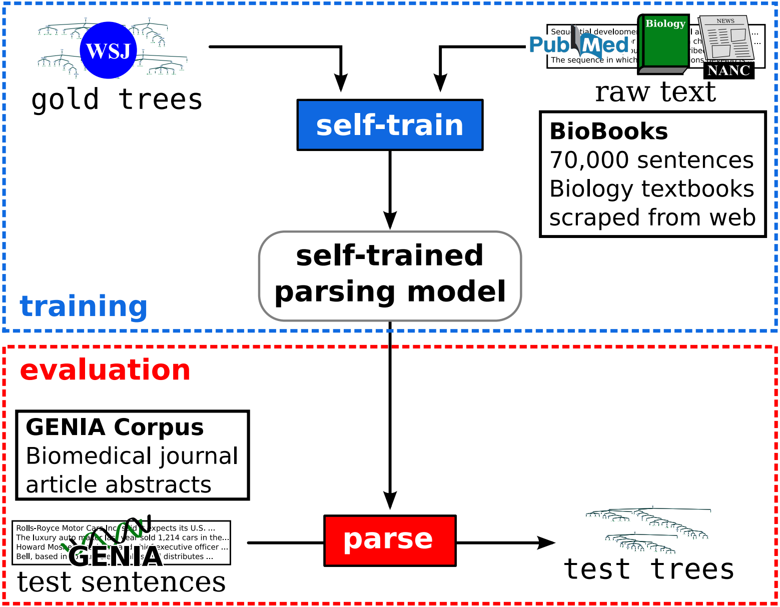
Parser adaptation experiments

More distant domains...



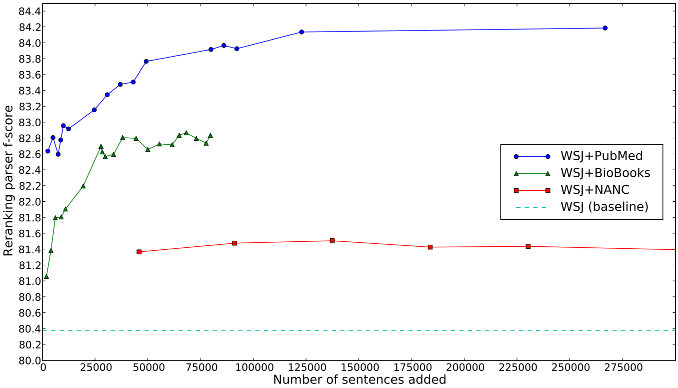
Parser adaptation experiments

More distant domains...



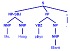
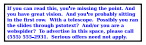
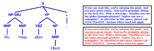
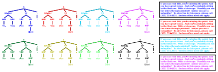
Varying unlabeled data for self-training

[McClosky and Charniak, ACL 2008]



f-score on GENIA development section

Parsing training scenarios

Train	Test	Description
		supervised parsing
		semi-supervised parsing
		parser portability
		parser adaptation
		multiple source parser adaptation

Automatic Domain Adaptation

- What if we don't know the target domain?

Automatic Domain Adaptation

- What if we don’t know the target domain?
    - Parsing the web or any other large heterogeneous corpus

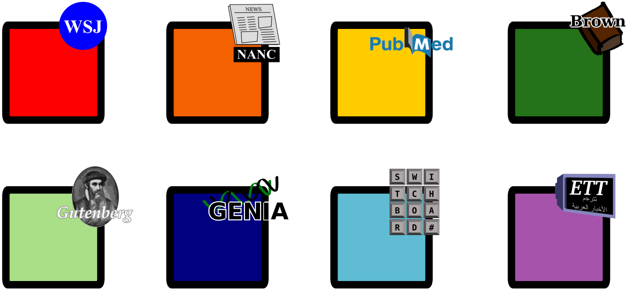
Automatic Domain Adaptation

- What if we don't know the target domain?
  - Parsing the web or any other large heterogeneous corpus
- Consider a new parsing task:
  - labeled and unlabeled corpora (source domains)

Automatic Domain Adaptation

- What if we don't know the target domain?
  - Parsing the web or any other large heterogeneous corpus
- Consider a new parsing task:
  - labeled and unlabeled corpora (source domains)
  - documents to parse (target text)

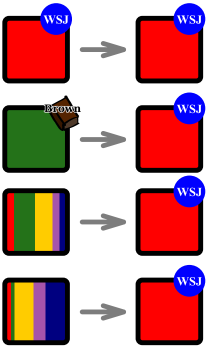
Any Domain Parsing



source domain models

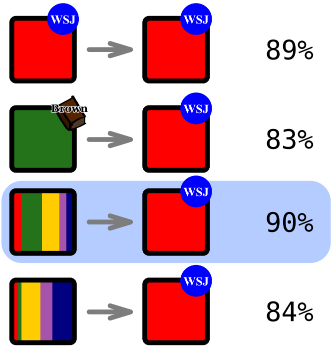
target text

Crossdomain Accuracy Prediction

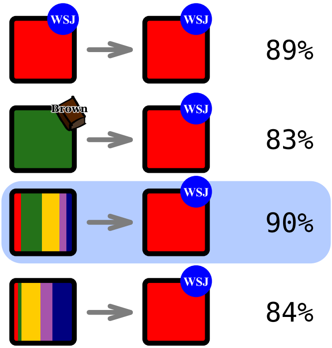


Crossdomain Accuracy Prediction

Crossdomain Accuracy Prediction



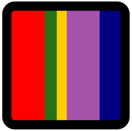
Crossdomain Accuracy Prediction



predict(, ?) = *f*-score

- Similar to [Ravi *et al.*, 2008]

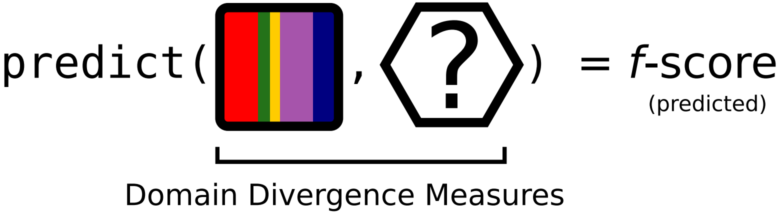
Prediction by regression

`predict(`  `,` 

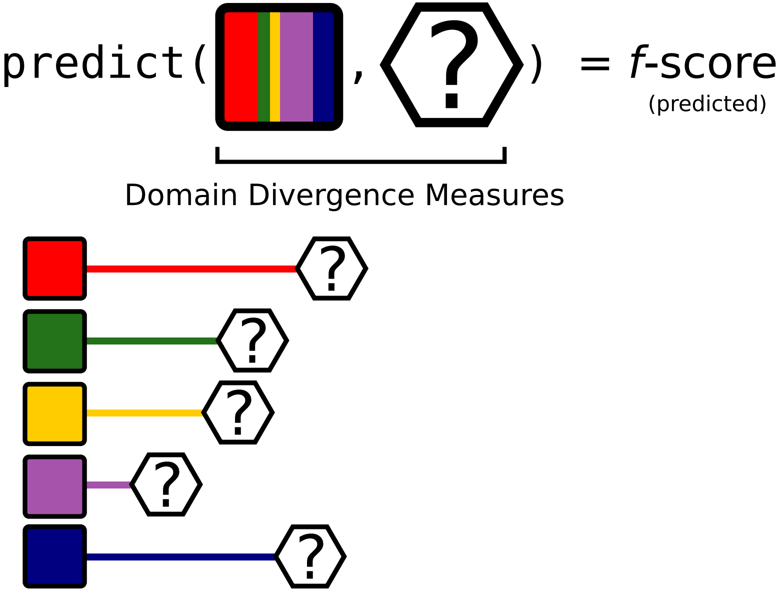
`)` = *f*-score

(predicted)

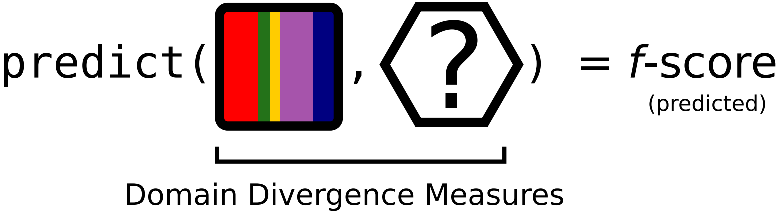
Regression features

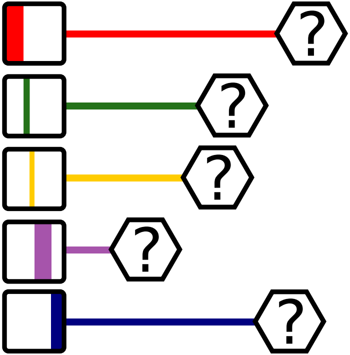


Regression features

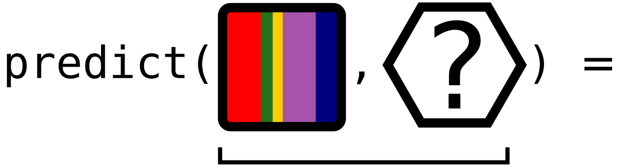


Regression features

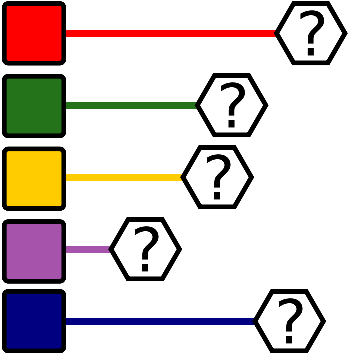




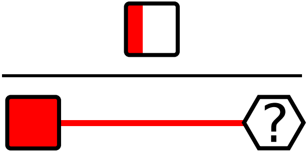
Regression features



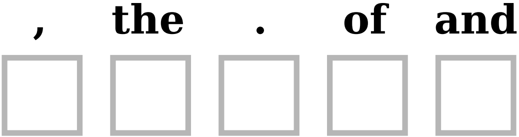
Domain Divergence Measures



Divide mixture weight by divergence:



Cosine Similarity



Cosine Similarity

	,	the	.	of	and
WSJ	4.9%	5.1%	3.8%	2.4%	1.8%
GENIA	3.6%	4.6%	3.6%	4.2%	2.6%

Cosine Similarity

	,	the	.	of	and
WSJ	4.9%	5.1%	3.8%	2.4%	1.8%
GENIA	3.6%	4.6%	3.6%	4.2%	2.6%

$$\text{cosine similarity} = \frac{\text{WSJ} \cdot \text{GENIA}}{\lVert \text{WSJ} \rVert \, \lVert \text{GENIA} \rVert} \approx 0.956$$

Unknown words

WSJ

Unknown words

WSJ

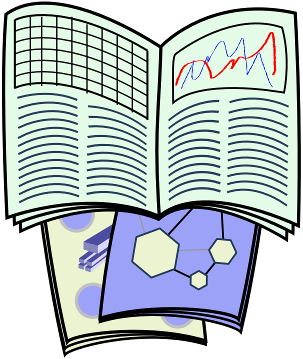
= vocabulary

Unknown words

WSJ

GENIA

= vocabulary

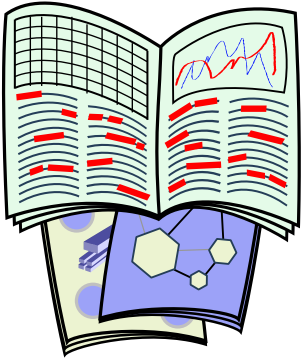


Unknown words

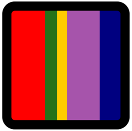
WSJ

GENIA

= vocabulary



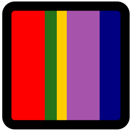
Regression features

predict( , ) = *f*-score

(predicted)

Domain Divergence Measures

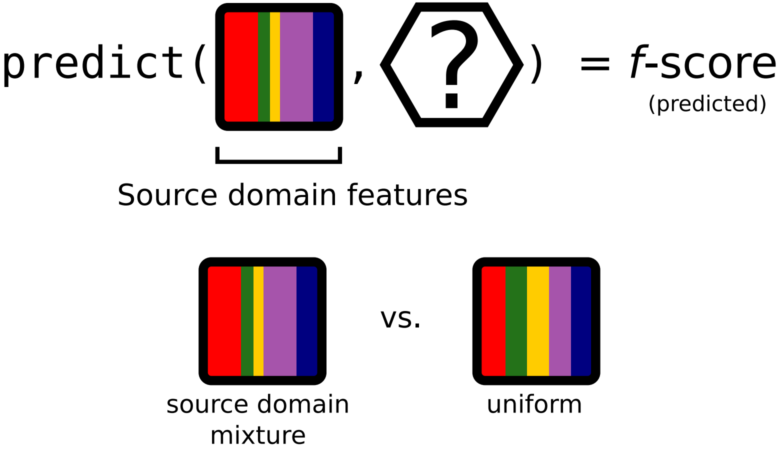
Regression features

predict( , ) = *f*-score

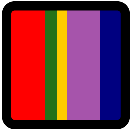
(predicted)

Source domain features

Regression features

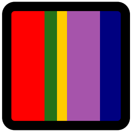


Regression features

predict( , ) = *f*-score

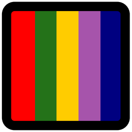
(predicted)

Source domain features



source domain mixture

vs.

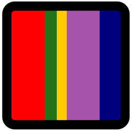


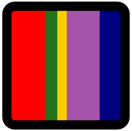
uniform

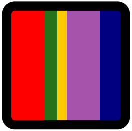
Entropy:

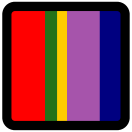
$$\mathrm{H}(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)$$

Model and estimation

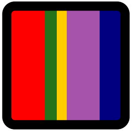
predict(, ) =

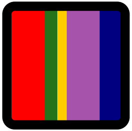
$\vec{\lambda_c}$ cosinesim(, )

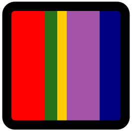
$+ \vec{\lambda_w}$ unkwords(, )

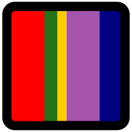
$+ \lambda_u$ entropy() + b

Model and estimation

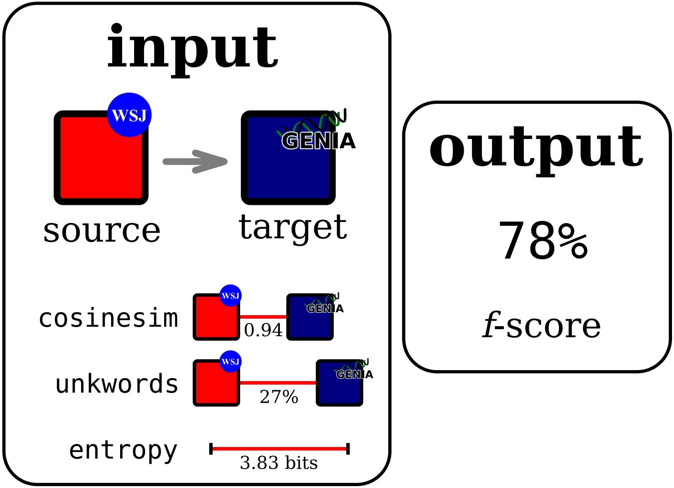
predict(, ) =

$\vec{\lambda}_c$ cosinesim(, )

$+ \vec{\lambda}_w$ unkwords(, )

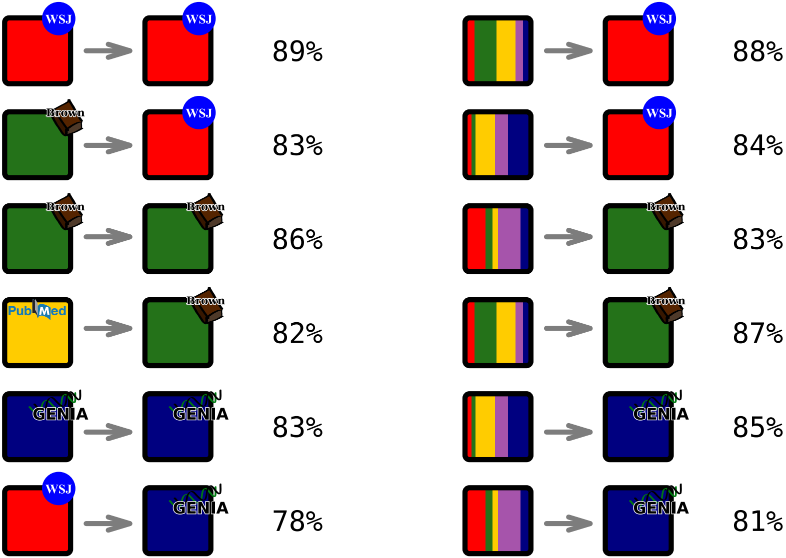
$+ \lambda_u$ entropy() $+$ b

Anatomy of a data point



* numbers on this slide are cooked

Training data



* numbers on this slide are cooked

Round-robin evaluation

WSJ

Brown

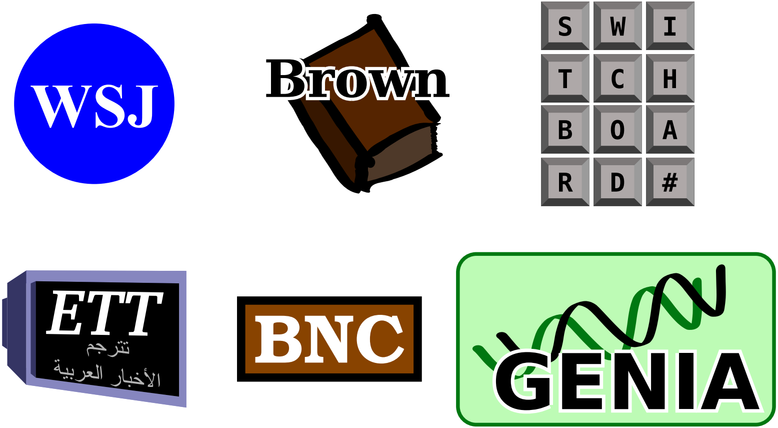
S W I
T C H
B O A
R D #

ETT
ترجم
الأخبار العربية

BNC

GENIA

Round-robin evaluation



Evaluation for GENIA

train

sources

WSJ Brown ETT
NANC Gutenberg SWITCHBOARD

targets

WSJ Brown ETT
SWITCHBOARD BNC

Evaluation for GENIA

train

sources

WSJ Brown ETT
NANC Gutenberg SWITCHBOARD

targets

WSJ Brown ETT
SWITCHBOARD BNC

test

sources

WSJ Brown ETT
NANC Gutenberg SWITCHBOARD

target

GENIA

Baselines

- Standard baselines

Baselines

- Standard baselines
  - Fixed set: WSJ

Baselines

- Standard baselines
  - Fixed set: WSJ
  - Uniform (no self-trained corpora)

Baselines

- Standard baselines
  - Fixed set: WSJ
  - Uniform (no self-trained corpora)
  - Uniform (all corpora)

Baselines

- Standard baselines
  - Fixed set: WSJ
  - Uniform (no self-trained corpora)
  - Uniform (all corpora)
- Oracle baselines

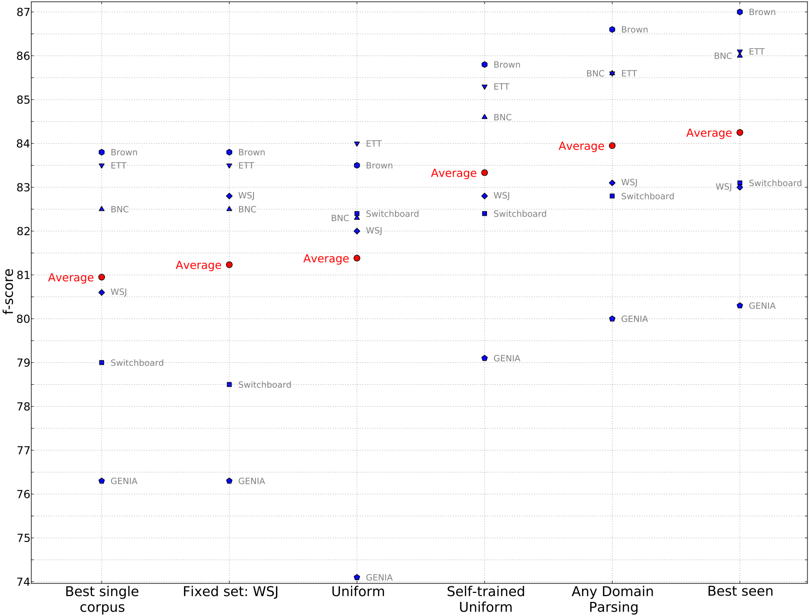
Baselines

- Standard baselines
  - Fixed set: WSJ
  - Uniform (no self-trained corpora)
  - Uniform (all corpora)
- Oracle baselines
  - Best single corpus

Baselines

- Standard baselines
  - Fixed set: WSJ
  - Uniform (no self-trained corpora)
  - Uniform (all corpora)
- Oracle baselines
  - Best single corpus
  - Best seen

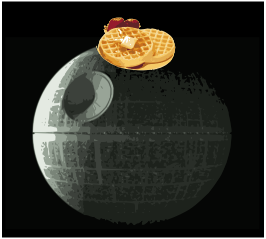
Evaluation results



Moral of the story

- What's the best way to parse new text?
  - Self-training on similar text improves performance
  - Any Domain Parsing provides additional benefits by selecting relevant corpora
- Self-training helps in many different parsing scenarios
  - Allows us to use unlabeled data to improve performance
  - State-of-the-art performance on WSJ, BROWN, and GENIA
- Relevant publications:
  - [McClosky, Charniak, and Johnson, NAACL 2006]
  - [McClosky, Charniak, and Johnson, ACL 2006]
  - [McClosky and Charniak, ACL 2008]
  - [McClosky, Charniak, and Johnson, COLING 2008]

Questions?



Thanks to my committee, BLLIP, friends, and family for their feedback and support!

Dedicated to my grandparents

Brought to you by NSF grants LIS9720368 and IIS0095940 and DARPA GALE contract HR0011-06-2-0001

Extra slides

What is...the matrix?

	Test				
Train	Literature	BioMed	Phone	ETT	News
Literature	86.7	73.5	77.6	80.8	79.9
BioMed	65.7	84.6	50.5	67.1	64.6
Phone	75.8	63.6	88.2	76.2	69.8
ETT	76.2	65.7	74.5	82.4	72.6
News	84.1	76.2	76.7	82.2	89.7

(f-scores on all sentences in test sets)

The Four Hypotheses

[McClosky, Charniak, and Johnson, COLING 2008]

Four hypotheses:

1. Self-training works after a phase transition.
2. Self-trained parser makes fewer search errors.
3. Certain classes of reranker features benefit self-training.
4. Self-training teaches the parser about bilexical dependencies.

In-domain evaluation

train

sources

WSJ, Brown, ETT, GENIA, NANC, Gutenberg, SWITCHBOARD, PubMed

targets

WSJ, Brown, ETT, SWITCHBOARD, BNC

---

test

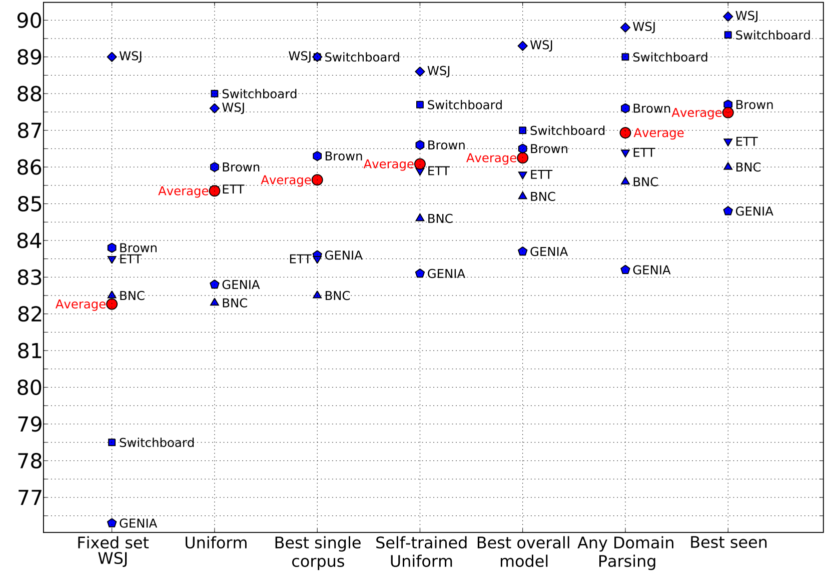
sources

WSJ, Brown, ETT, GENIA, NANC, Gutenberg, SWITCHBOARD, PubMed

target

GENIA

In-domain evaluation results



Analysis

[McClosky, Charniak, and Johnson, NAACL 2006, ACL 2006, COLING 2008]

- Self-trained parser is more confident.

Analysis

[McClosky, Charniak, and Johnson, NAACL 2006, ACL 2006, COLING 2008]

- Self-trained parser is more confident.
- Self-trained first stage parser has better potential.

Analysis

[McClosky, Charniak, and Johnson, NAACL 2006, ACL 2006, COLING 2008]

- Self-trained parser is more confident.
- Self-trained first stage parser has better potential.
- Factor analysis: predict when self-training might help
  - Sentence length

Analysis

[McClosky, Charniak, and Johnson, NAACL 2006, ACL 2006, COLING 2008]

- Self-trained parser is more confident.
- Self-trained first stage parser has better potential.
- Factor analysis: predict when self-training might help
  - Sentence length
  - # of conjunctions

Analysis

[McClosky, Charniak, and Johnson, NAACL 2006, ACL 2006, COLING 2008]

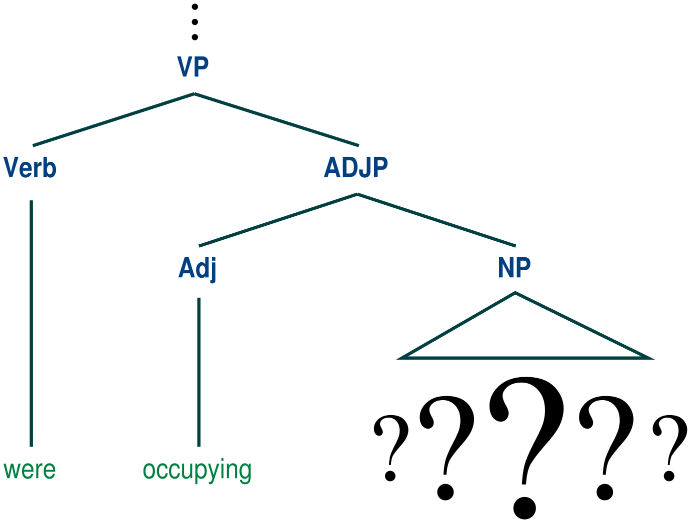
- Self-trained parser is more confident.
- Self-trained first stage parser has better potential.
- Factor analysis: predict when self-training might help
  - Sentence length
  - # of conjunctions
  - # of bigrams in NANC not seen in WSJ

Why does self-training help?

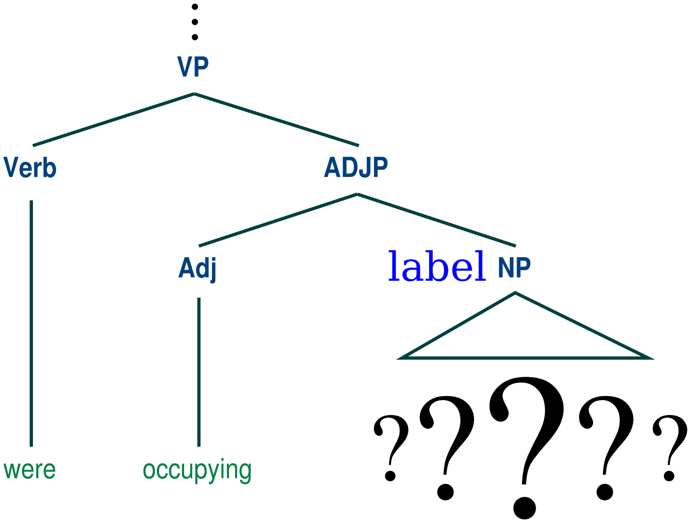
[McClosky, Charniak, and Johnson, NAACL 2006, ACL 2006, COLING 2008]

- Hypothesis:
  - Self-training teaches the parser about bilexical dependencies.

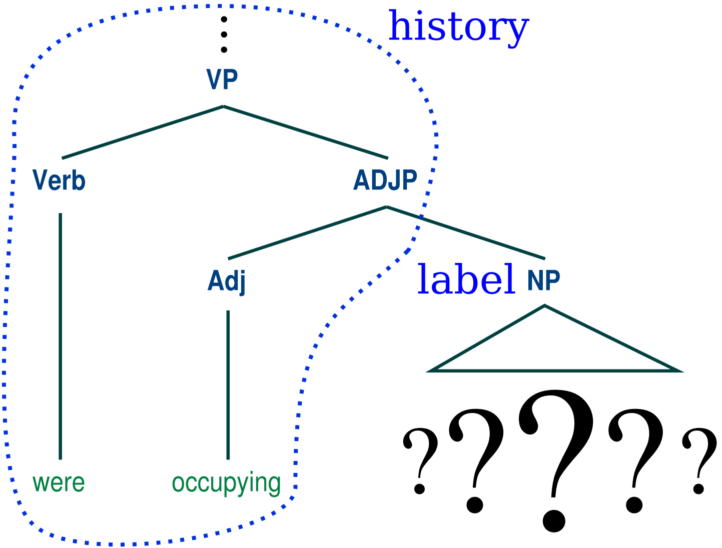
The Generative Story



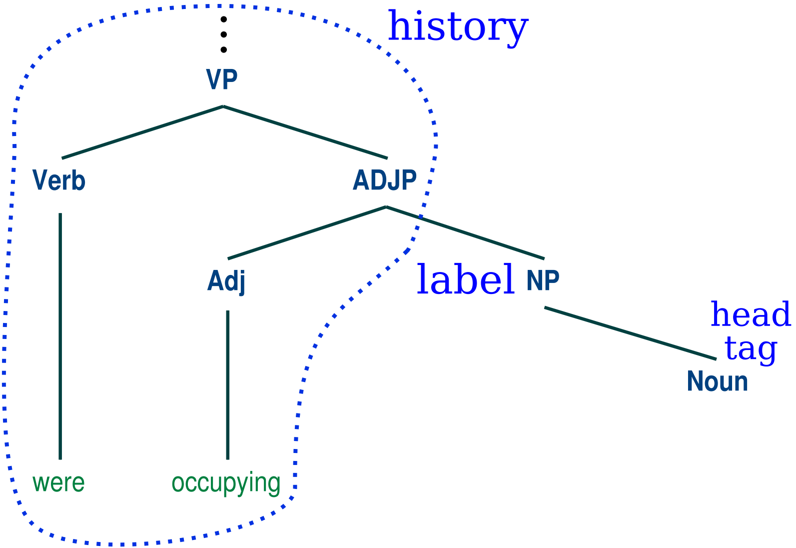
The Generative Story



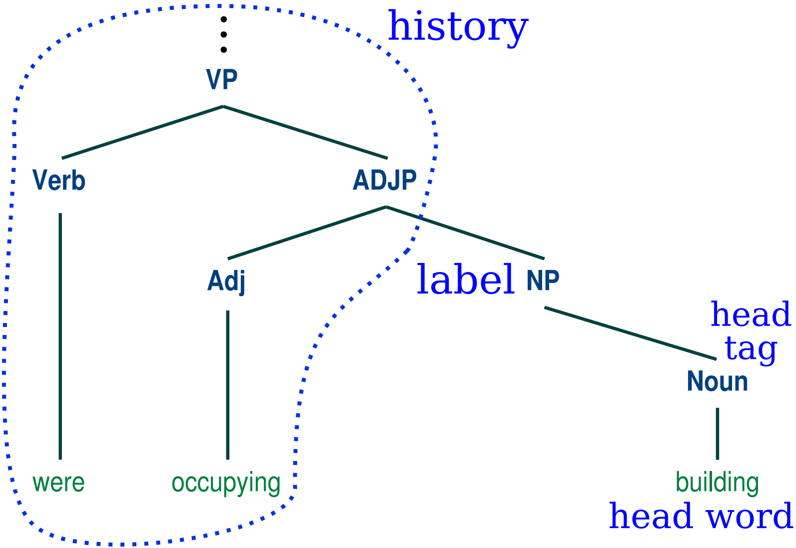
The Generative Story



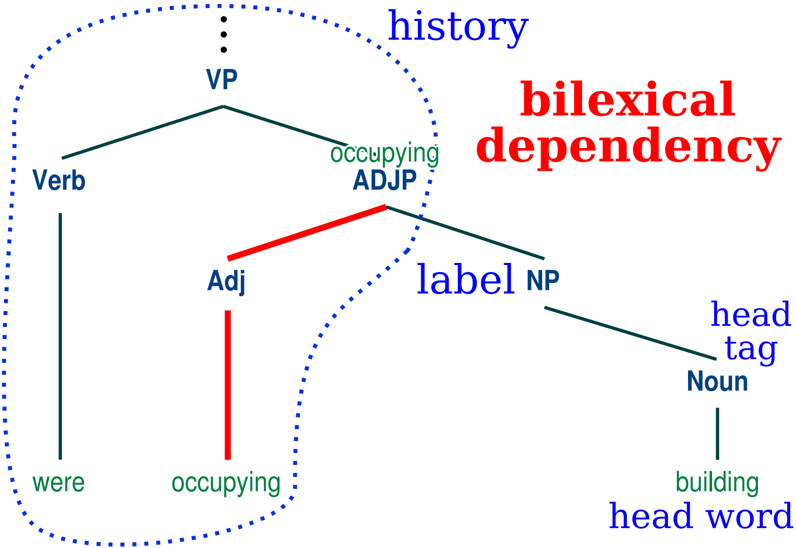
The Generative Story



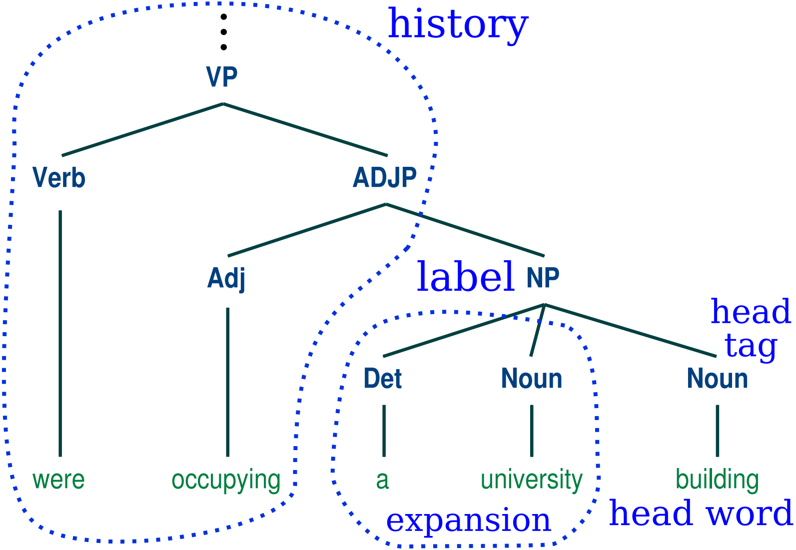
The Generative Story



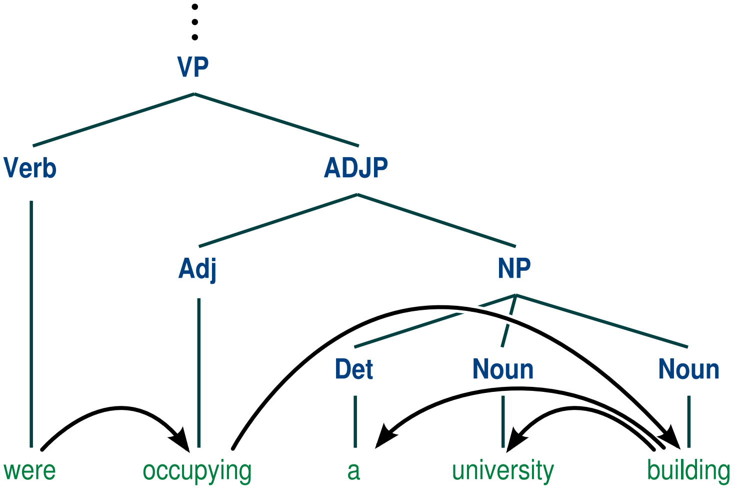
The Generative Story



The Generative Story



The Generative Story



What does self-training teach the parser?

[McClosky, Charniak, and Johnson, COLING 2008]

$$
\begin{aligned}
P(\mathit{constit} \mid \mathit{label}, \mathit{history}) &= P(\mathit{tag} \mid \mathit{label}, \mathit{history}) \\
&\times P(\mathit{head} \mid \mathit{tag}, \mathit{label}, \mathit{history}) \\
&\times P(\mathit{exp} \mid \mathit{head}, \mathit{tag}, \mathit{label}, \mathit{history})
\end{aligned}
$$

What does self-training teach the parser?

[McClosky, Charniak, and Johnson, COLING 2008]

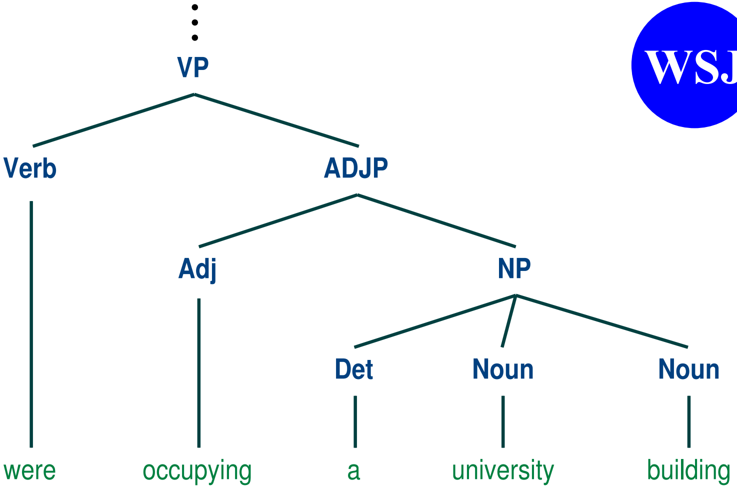
$$
\begin{aligned}
P(\mathit{constit} \mid \mathit{label}, \mathit{history}) &= P(\mathit{tag} \mid \mathit{label}, \mathit{history}) \\
&\times P(\mathit{head} \mid \mathit{tag}, \mathit{label}, \mathit{history}) \\
&\times P(\mathit{exp} \mid \mathit{head}, \mathit{tag}, \mathit{label}, \mathit{history})
\end{aligned}
$$

What does self-training teach the parser?

[McClosky, Charniak, and Johnson, COLING 2008]

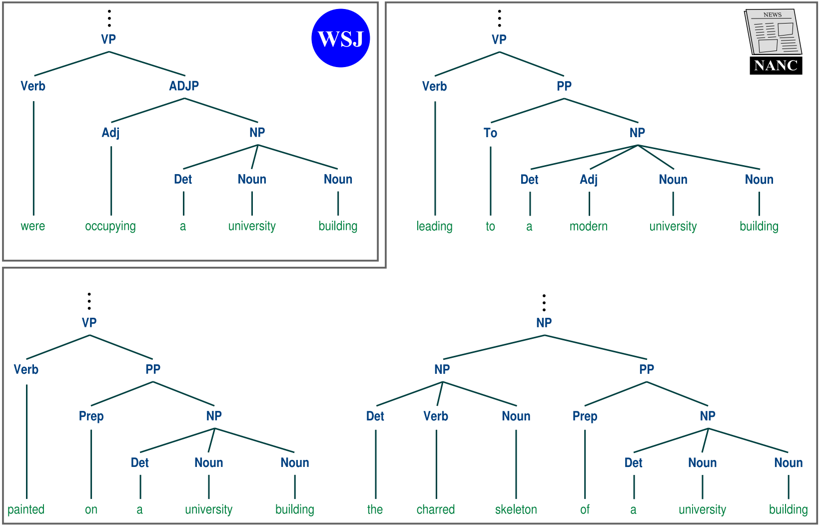
$$\begin{aligned} P(\mathit{constit} \mid \mathit{label}, \mathit{history}) &= P(\mathit{tag} \mid \mathit{label}, \mathit{history}) \\ &\times P(\mathit{head} \mid \mathit{tag}, \mathit{label}, \mathit{history}) \\ &\times P(\mathit{exp} \mid \mathit{head}, \mathit{tag}, \mathit{label}, \mathit{history}) \end{aligned}$$

What we learn from unlabeled data

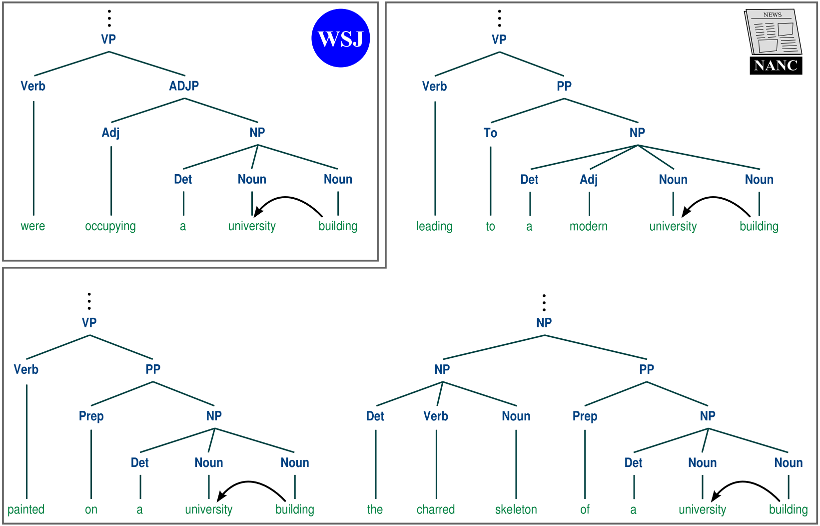


WSJ

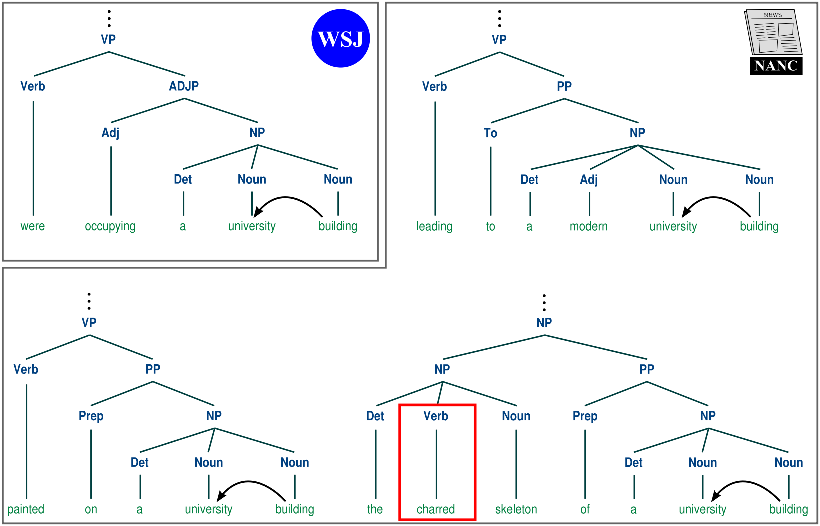
What we learn from unlabeled data



What we learn from unlabeled data



What we learn from unlabeled data



Outline

- Introduction
  - Parsing basics and applications
  - Domain Dependence
- Self-training for Parsing
  - Self-training
  - Semi-supervised parsing
  - Parser portability
  - Parser adaptation
- Any Domain Parsing: Automatic Domain Adaptation
  - Any Domain Parsing
  - Estimation and training data
  - Evaluation
  - Results
- Conclusion
- Extra slides
  - In-domain evaluation
  - Analysis