Domain Adaptation for Constituency Parsing Using Partial Annotations

Vidur Joshi
Matthew Peters
Mark Hopkins

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

Constituency Parsing is Useful

Textual Entailment (Bowman et al., 2016)

Semantic Parsing (Hopkins et al., 2017)

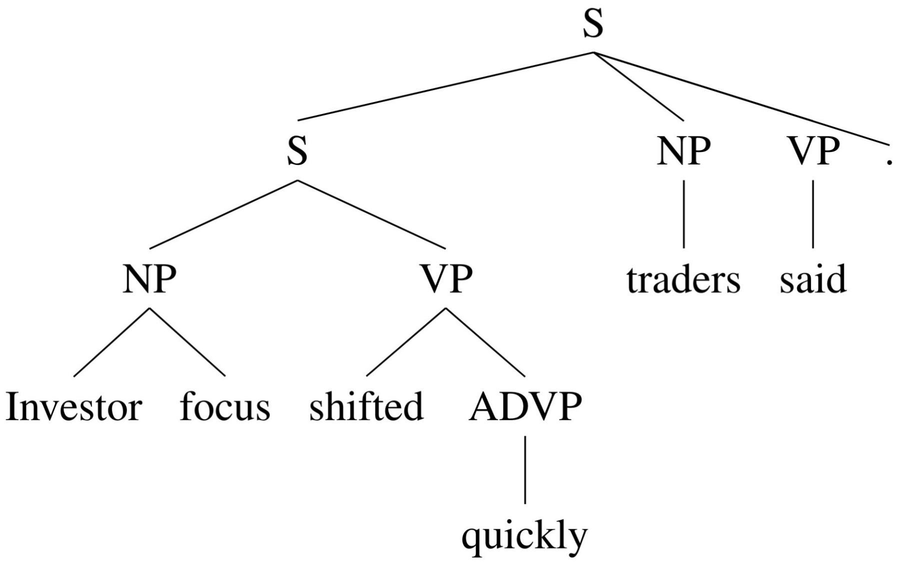
Sentiment Analysis (Socher et al., 2013)

Language Modeling (Dyer et al., 2016)

Penn Tree Bank (PTB) (Marcus et al., 1993)

40,000 annotated sentences

Newswire domain



But, Target Domains Are Diverse!

Geometry Problem:

In the rhombus PQRS, PR = 24 and QS = 10.

Question:

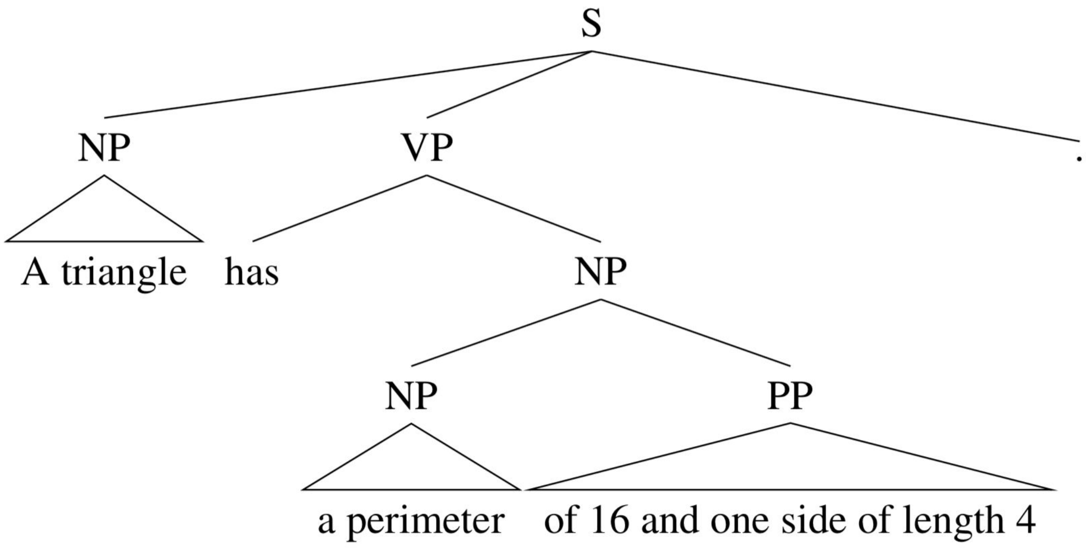
What's the second-most-used vowel in English?

Biochemistry:

Ethoxycoumarin was metabolized by isolated epidermal cells via dealkylation to 7-hydroxycoumarin (7-OHC) and subsequent conjugation.

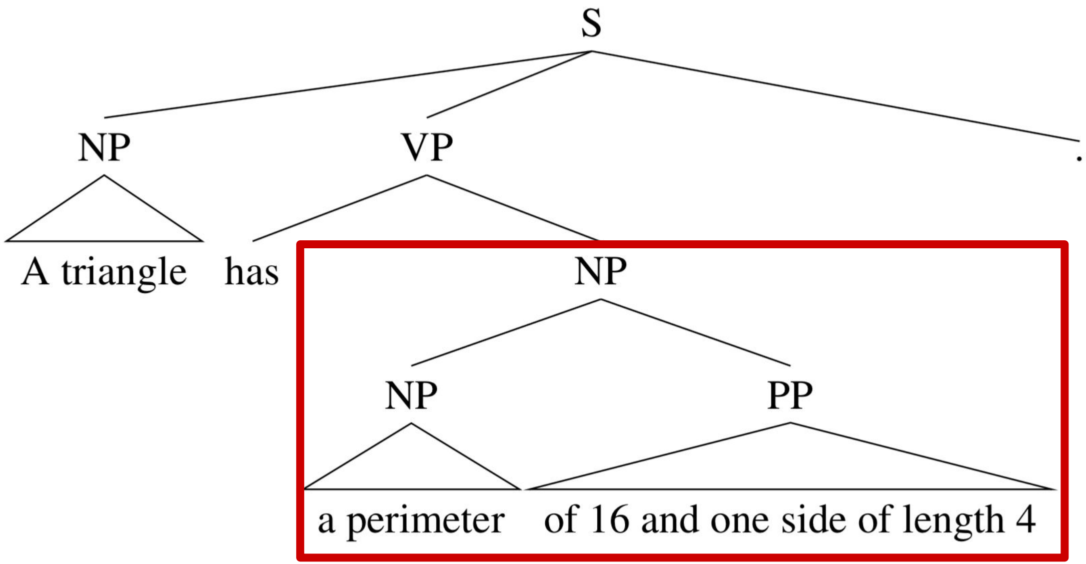
Performance Outside Source Domain

Parse geometry sentence with PTB trained parser



Performance Outside Source Domain

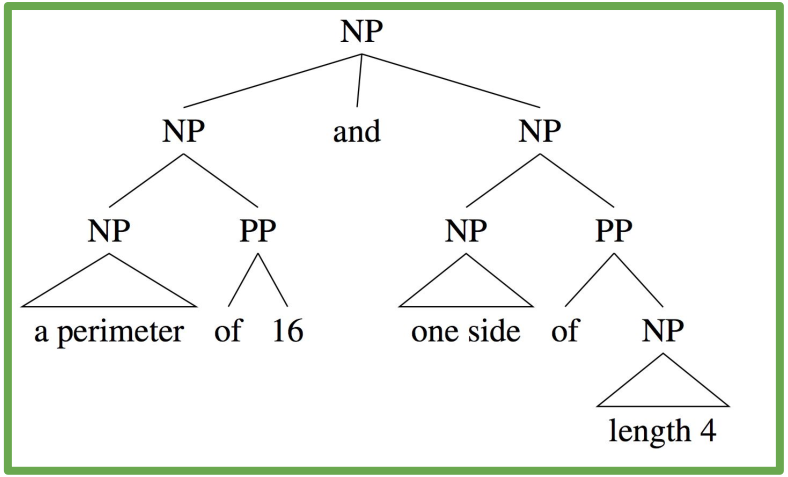
Parse geometry sentence with PTB trained parser

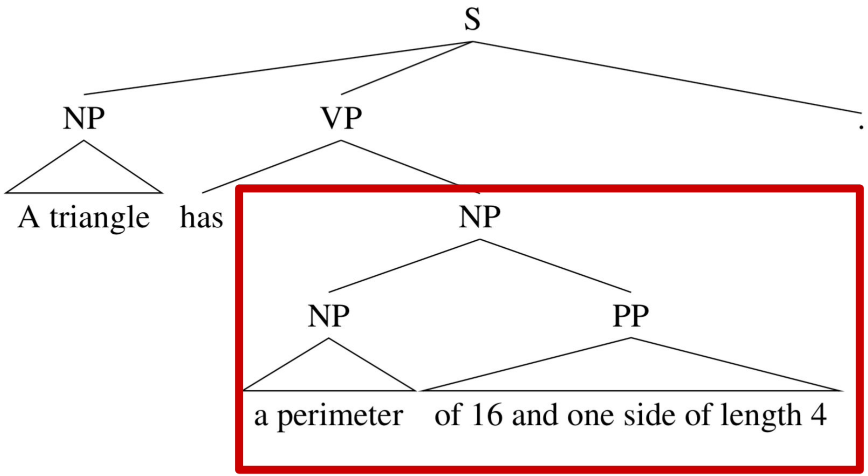


ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

Performance Outside Source Domain

Parse geometry sentence with PTB trained parser





How can we cheaply create high quality parsers for new domains?

Relevant Recent Developments in NLP

Contextualized word representations improve sample efficiency. (Peters et al., 2018)

Span-focused models achieve state-of-the-art constituency parsing results. (Stern et al., 2017)

Contributions

Show contextual word embeddings help domain adaptation.
E.g., **Over 90% F1 on Brown Corpus.**

Adapt a parser using partial annotations.
E.g., **Increase correct geometry-domain parses by 23%.**

Outline

Review Contextual Word Representations

Partial Annotations:

- Definition
- Training
- Parsing as Span Classification
- The Span Classification Model

Experiments and Results:

- Performance on PTB and new Domains
- Adapting Using Partial Annotations

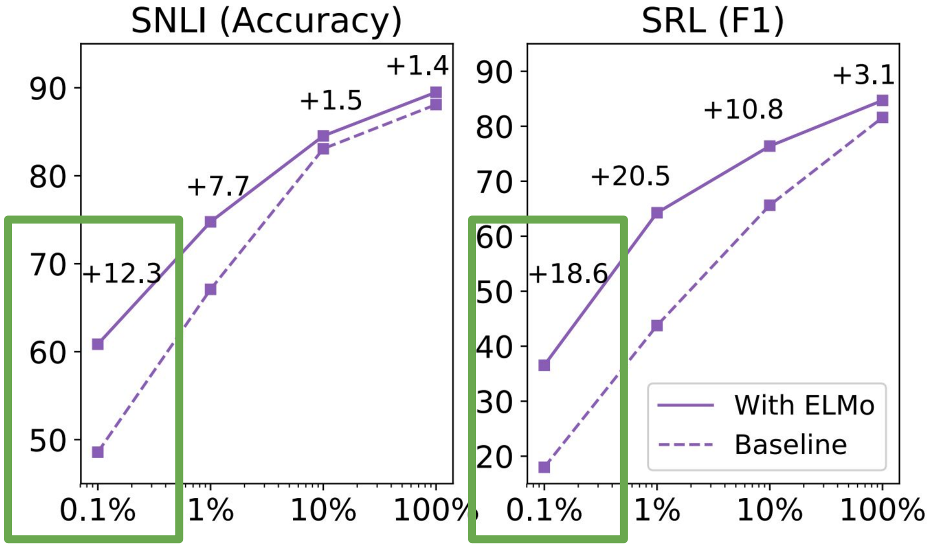
Contextualized Word Representations

ELMo trained on Billion Word Corpus (Peters et al., 2018).

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

Contextualized Word Representations

ELMo trained on Billion Word Corpus (Peters et al., 2018).



Improve sample efficiency

Partial Annotations

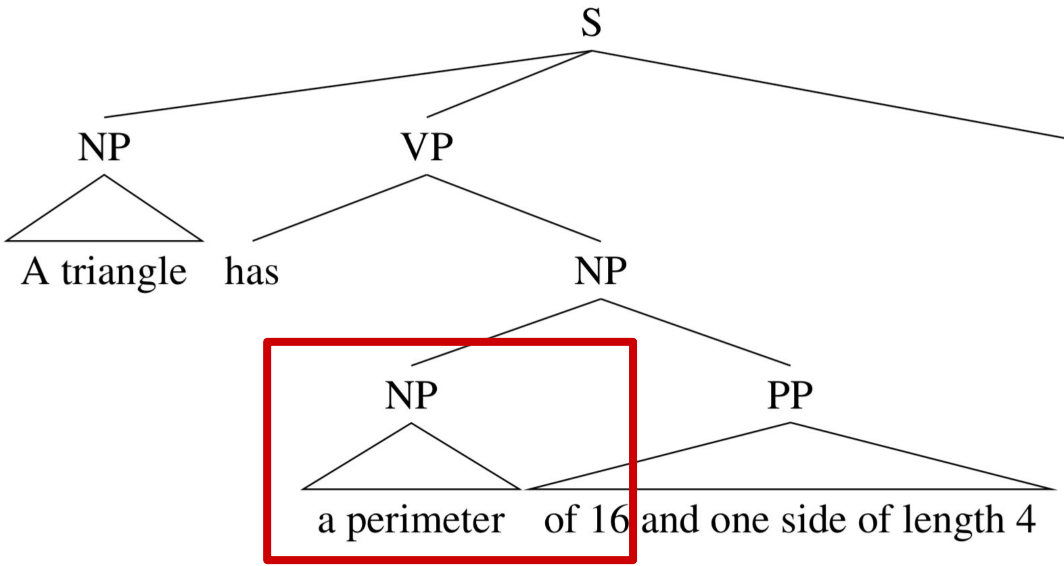
Definition

Training

Parsing as Span Classification

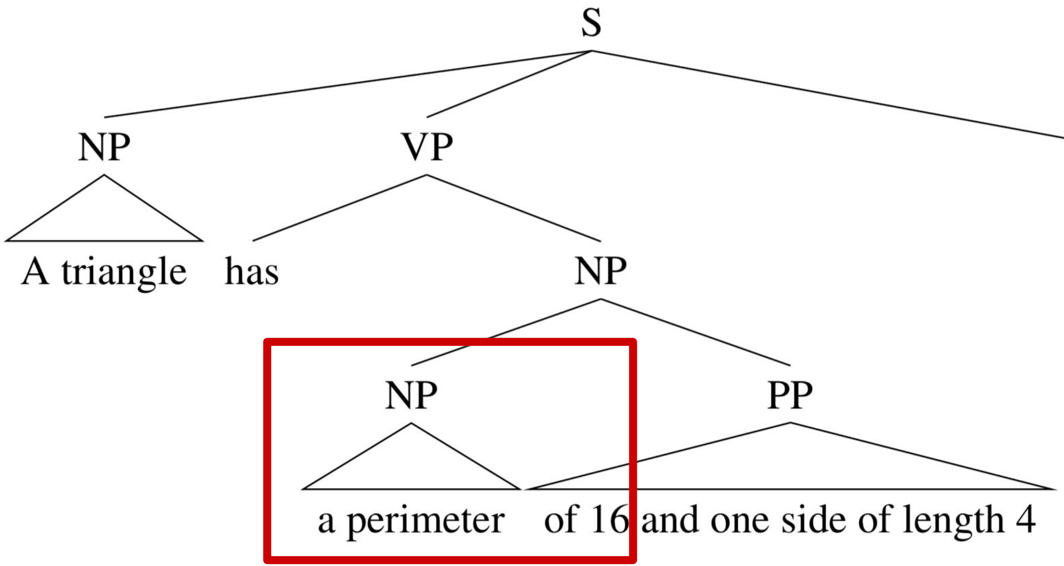
The Span Classification Model

Selectively Annotate Important Phenomena



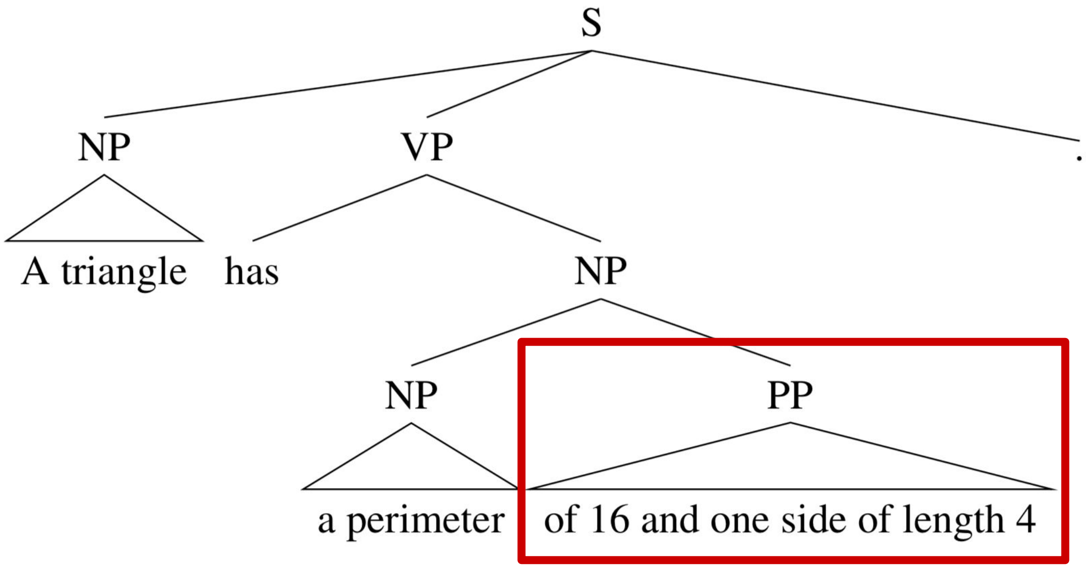
A triangle has a perimeter of 16 and one side of length 4.

Selectively Annotate Important Phenomena



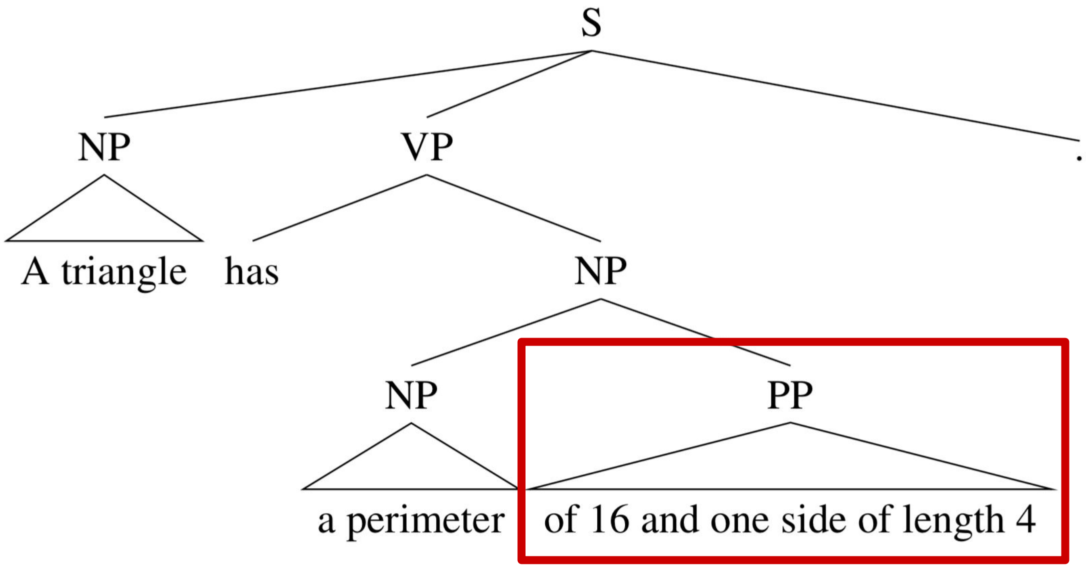
A triangle has [a perimeter of 16] and one side of length 4.

Selectively Annotate Important Phenomena



A triangle has [a perimeter of 16] and one side of length 4.

Selectively Annotate Important Phenomena



A triangle has [a perimeter {of 16] and one side of length 4}.

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

Full Versus Partial Annotation

(S (NP **A triangle**) (VP **has** (NP (NP (NP **a perimeter**) (PP **of 16**)) **and** (NP (NP **one side**) (PP **of** (NP **length 4**))))) **.**)

A triangle has [a perimeter {of 16] and one side of length 4}.

Partial Annotation Definition

Partial annotation is a labeled span.

A triangle has [a perimeter of 16] and one side of length 4 .

A triangle has [NP a perimeter of 16] and one side of length 4 .

A triangle has a perimeter {of 16 and one side of length 4} .

Why Partial Annotations?

Allowing annotators to selectively annotate important phenomena, makes the process faster and simpler.

(Mielens et al., 2015)

Definition

Training

Parsing as Span Classification

The Span Classification Model

Objective for Full Annotation

$$\mathcal{L}(\theta) = - \sum_{(\text{sentence},\text{parse})} \log \Pr_\theta(\text{parse}|\text{sentence})$$

Objective for Partial Annotation

Since we do not have a full parse,

marginalize out components for which no supervision exists.

$$\mathcal{L}(\theta) = - \sum_{(\text{sentence},\text{annotations})} \log \left(\boxed{\sum_{\text{parses consistent with annotations}} \Pr_\theta(\text{parse}|\text{sentence})} \right)$$

Objective for Partial Annotation

Marginalize out components for which no supervision exists.

$$\mathcal{L}(\theta) = - \sum_{(\text{sentence},\text{annotations})} \log \left(\boxed{\sum_{\text{parses consistent with annotations}}} \Pr_\theta(\text{parse} | \text{sentence}) \right)$$

Expensive!

ALLENAI.ORG ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

One Solution: Approximation*

$$\mathcal{L}(\theta) = -\sum_{\text{(sentence,annotations)}} \log \left(\sum_{\boxed{\text{top k parses consistent with annotations}}} \Pr_\theta(\text{parse}|\text{sentence}) \right)$$

*(Mirroshandel and Nasr, 2011; Majidi and Crane, 2013, Nivre et al., 2014; Li et al., 2016)

Our Solution: Parsing as Span Classification

Assume probability of a parse factors into a product of probabilities.

$$\Pr_\theta(\text{parse}|\text{sentence}) = \prod_{(\text{span},\text{label}) \text{ consistent with parse}} \Pr_\theta(\text{label}|\text{sentence}, \text{span})$$

Our Solution: Parsing as Span Classification

Assume probability of a parse factors into a product of probabilities.

$$\Pr_\theta(\text{parse}|\text{sentence}) = \boxed{\prod_{(\text{span},\text{label})\ \text{consistent with parse}}} \Pr_\theta(\text{label}|\text{sentence}, \text{span})$$

Our Solution: Parsing as Span Classification

Assume probability of a parse factors into a product of probabilities.

$$\Pr_\theta(\text{parse}|\text{sentence}) = \prod_{(\text{span,label})\ \text{consistent with parse}} \boxed{\Pr_\theta(\text{label}|\text{sentence, span})}$$

Our Solution: Parsing as Span Classification

Assume probability of a parse factors into a product of probabilities.

$$\Pr_\theta(\text{parse}|\text{sentence}) = \prod_{(\text{span},\text{label})\text{ consistent with parse}} \Pr_\theta(\text{label}|\text{sentence}, \text{span})$$

Objective now simplifies to:

$$\mathcal{L}(\theta) = -\sum_{(\text{sentence},\text{annotations})} \sum_{(\text{span},\text{label})\in\text{annotations}} \log \Pr_\theta(\text{label}|\text{sentence}, \text{span})$$

Easy if model classifies spans!

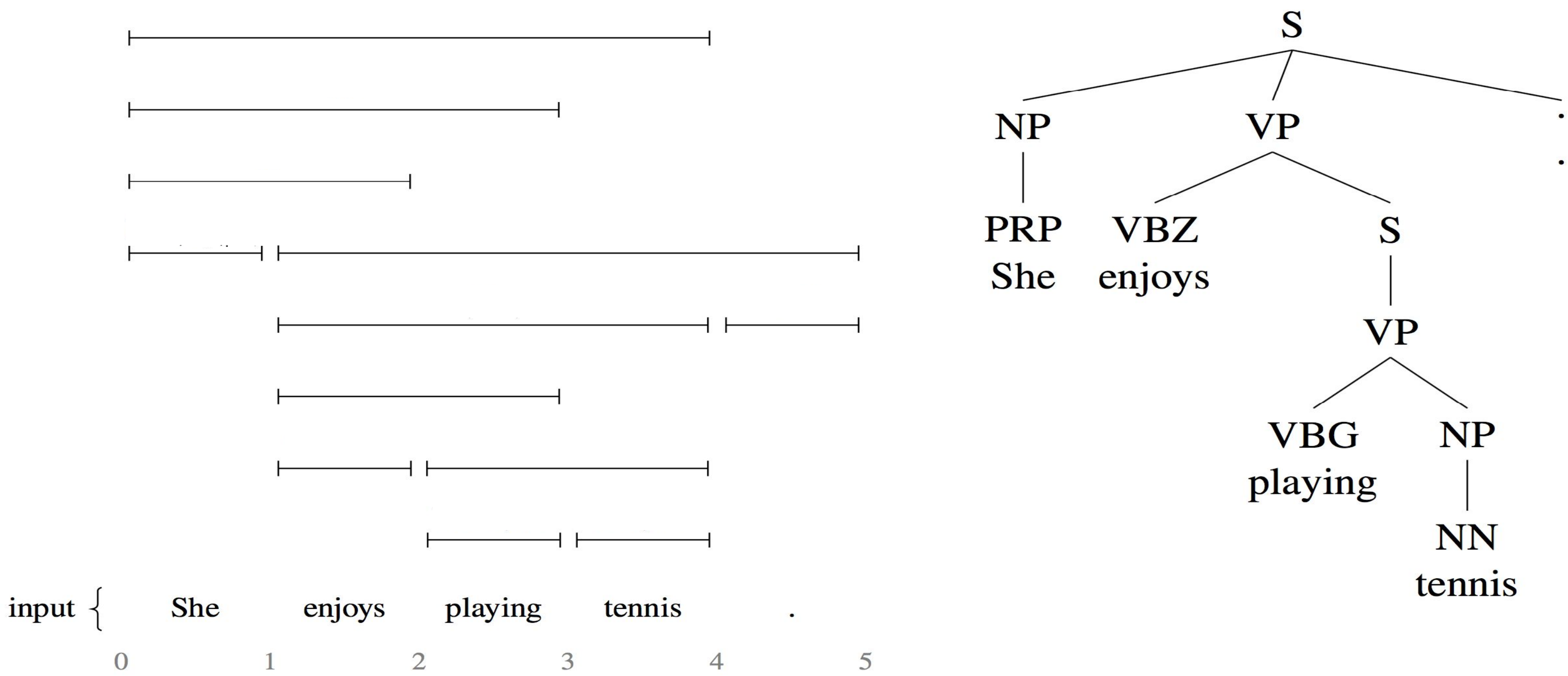
Definition

Training

Parsing as Span Classification

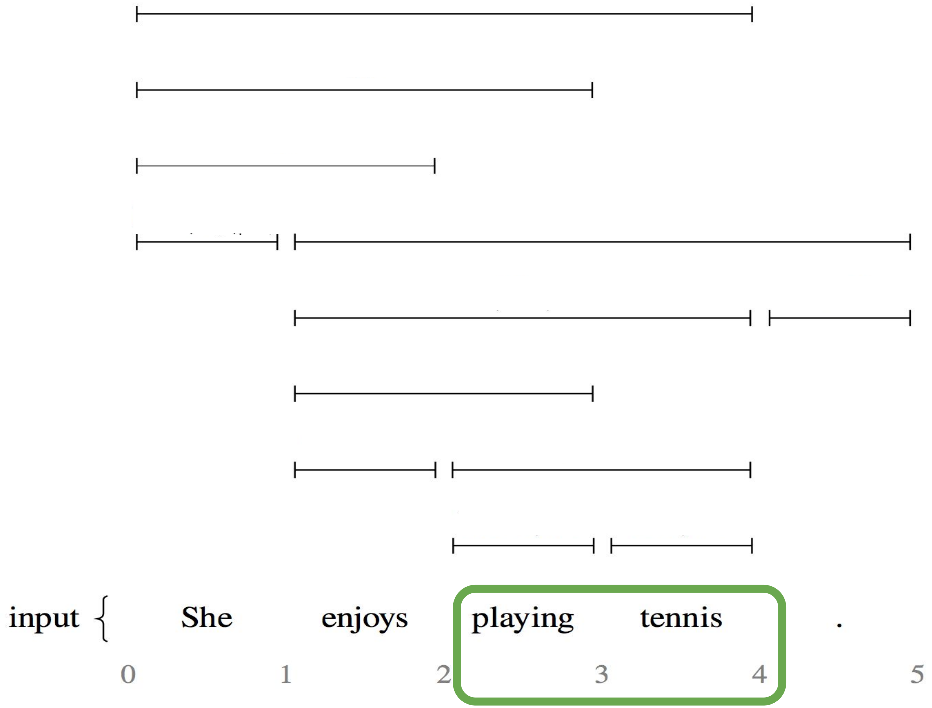
The Span Classification Model

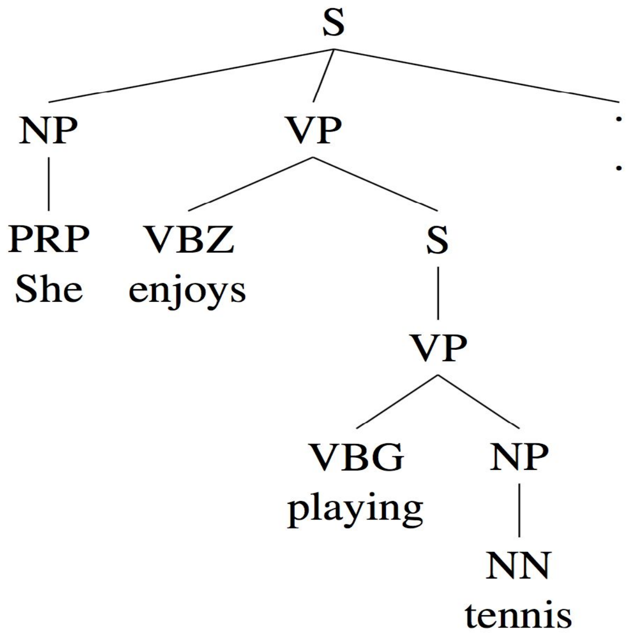
Parse Tree Labels All Spans*



*(Cross and Huang, 2016; Stern et al., 2017)

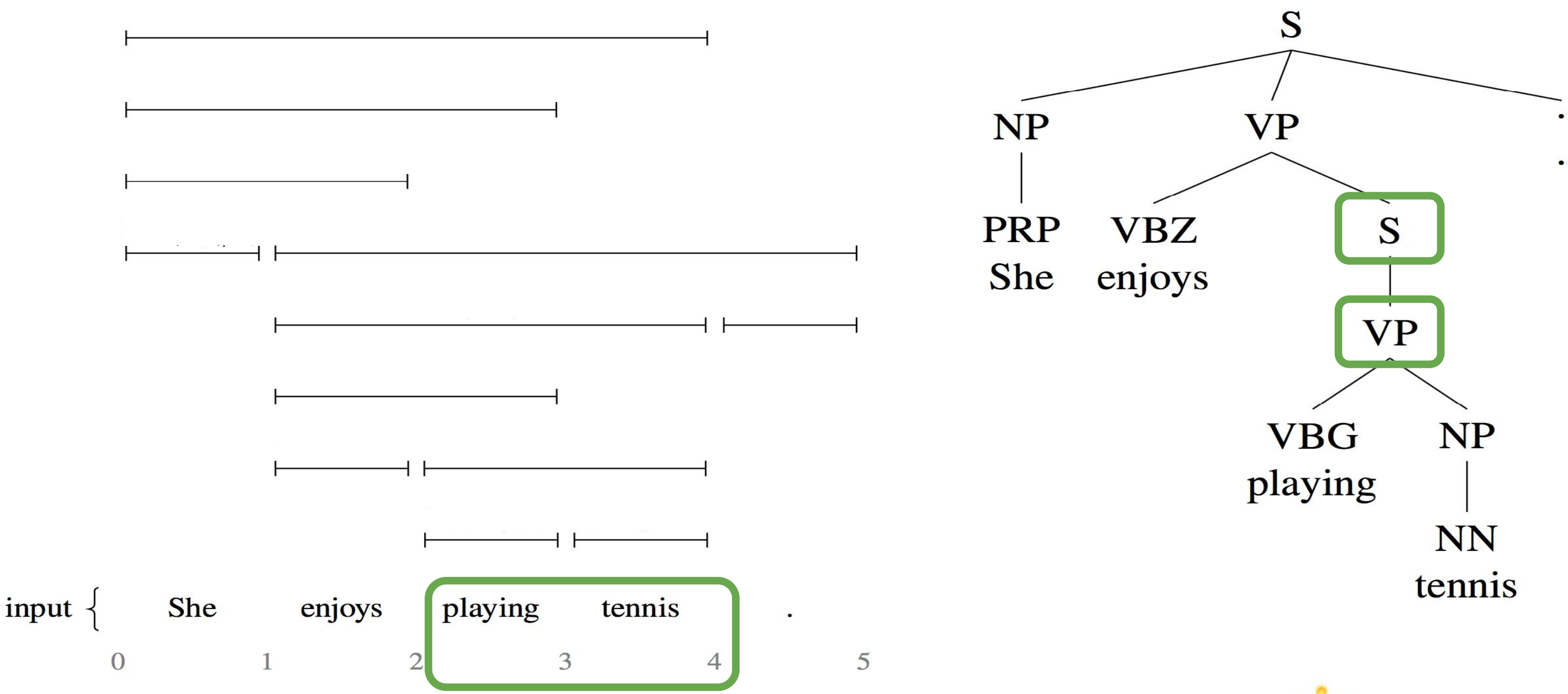
Parse Tree Labels All Spans*





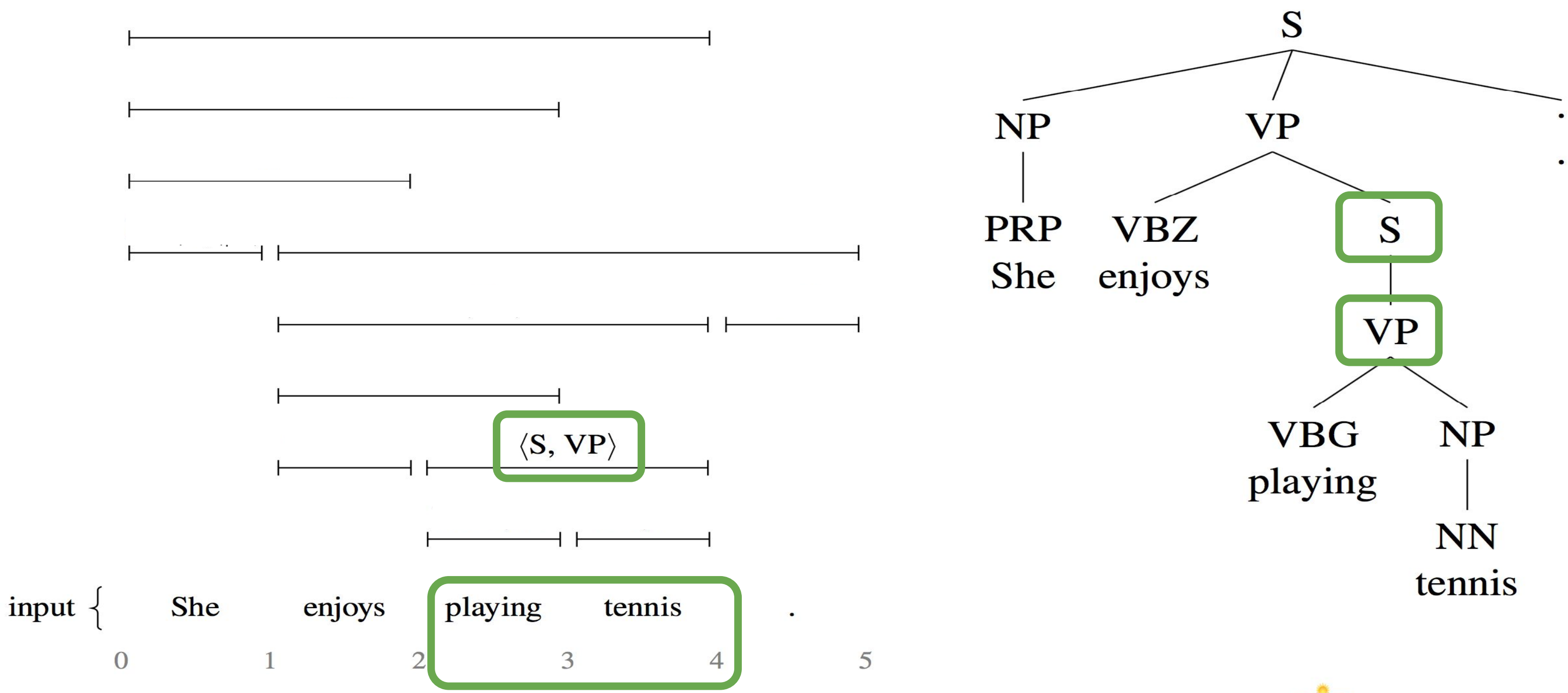
*(Cross and Huang, 2016; Stern et al., 2017)

Parse Tree Labels All Spans*



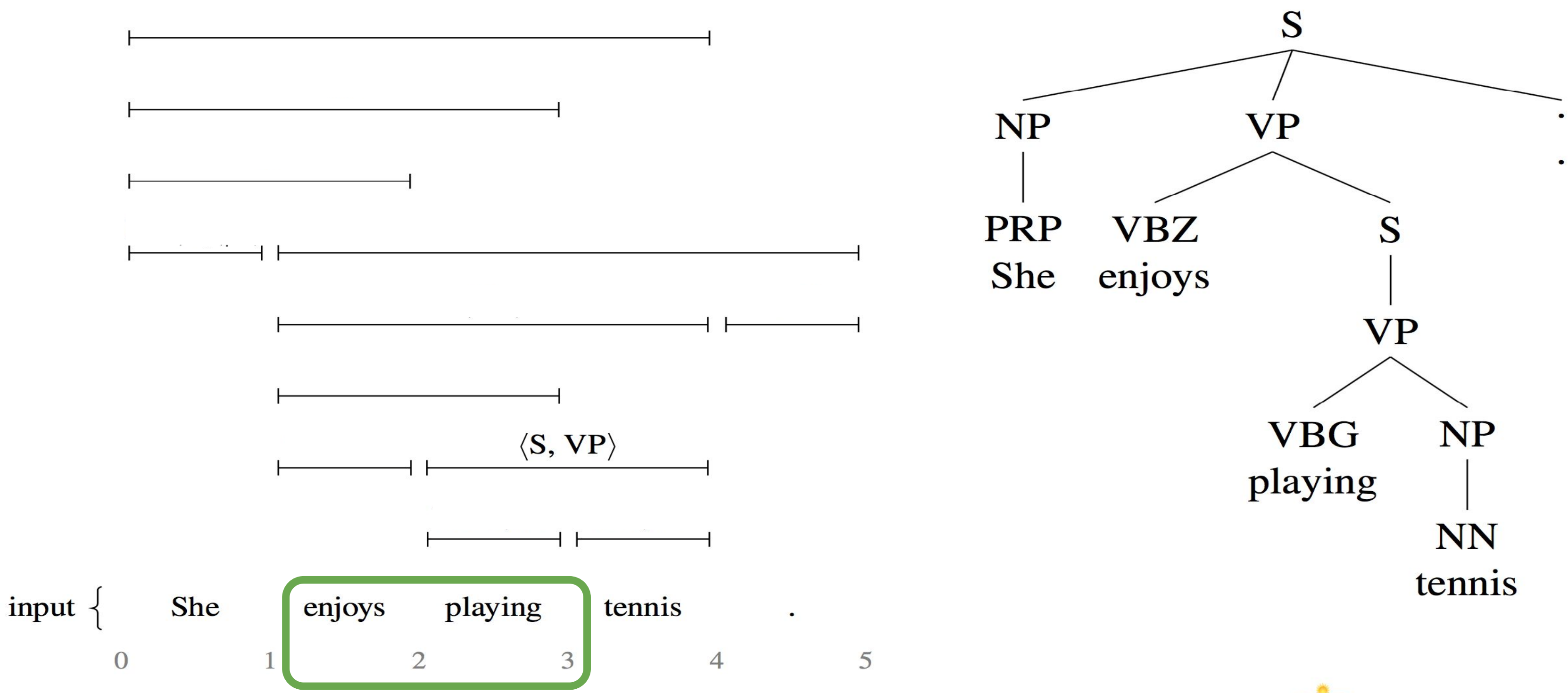
*(Cross and Huang, 2016; Stern et al., 2017)

Parse Tree Labels All Spans*



*(Cross and Huang, 2016; Stern et al., 2017)

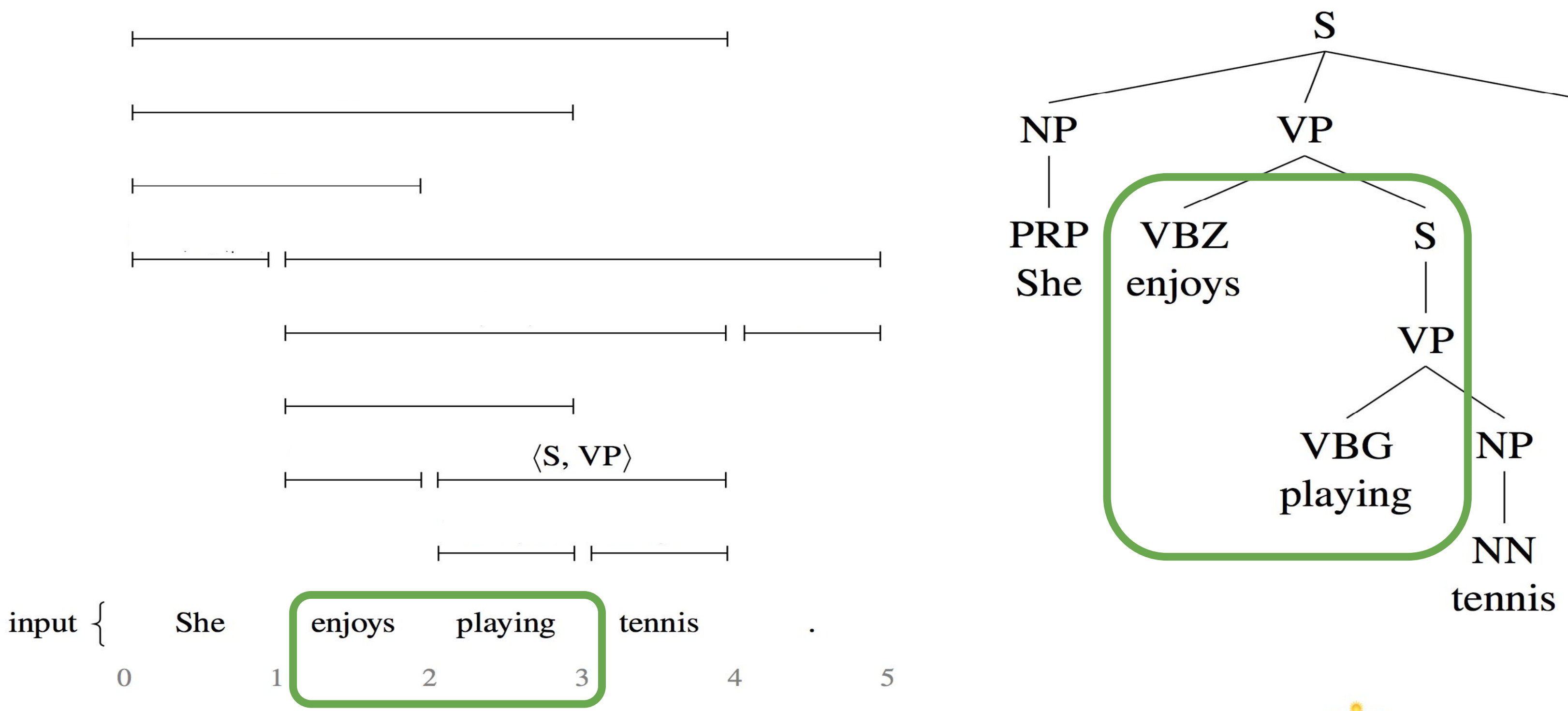
Parse Tree Labels All Spans*



*(Cross and Huang, 2016; Stern et al., 2017)

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

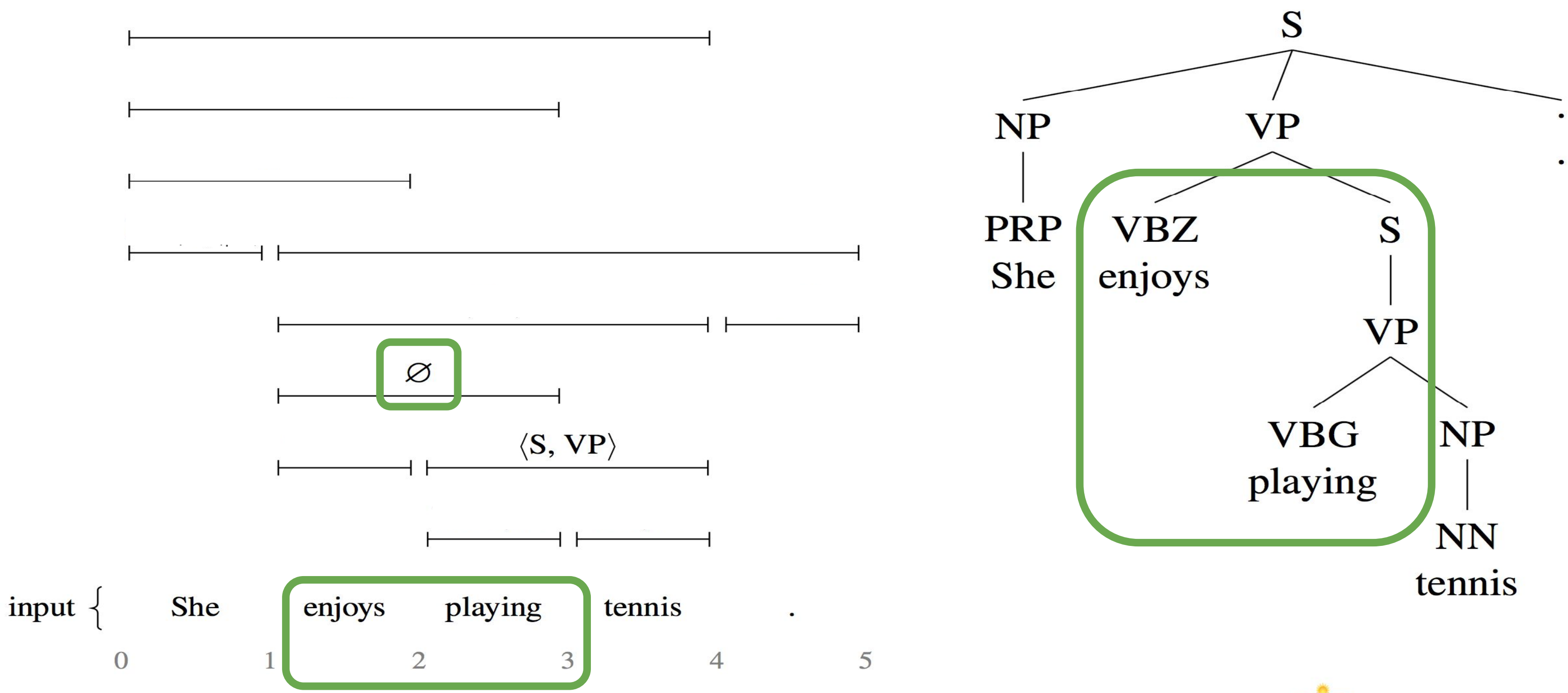
Parse Tree Labels All Spans*



*(Cross and Huang, 2016; Stern et al., 2017)

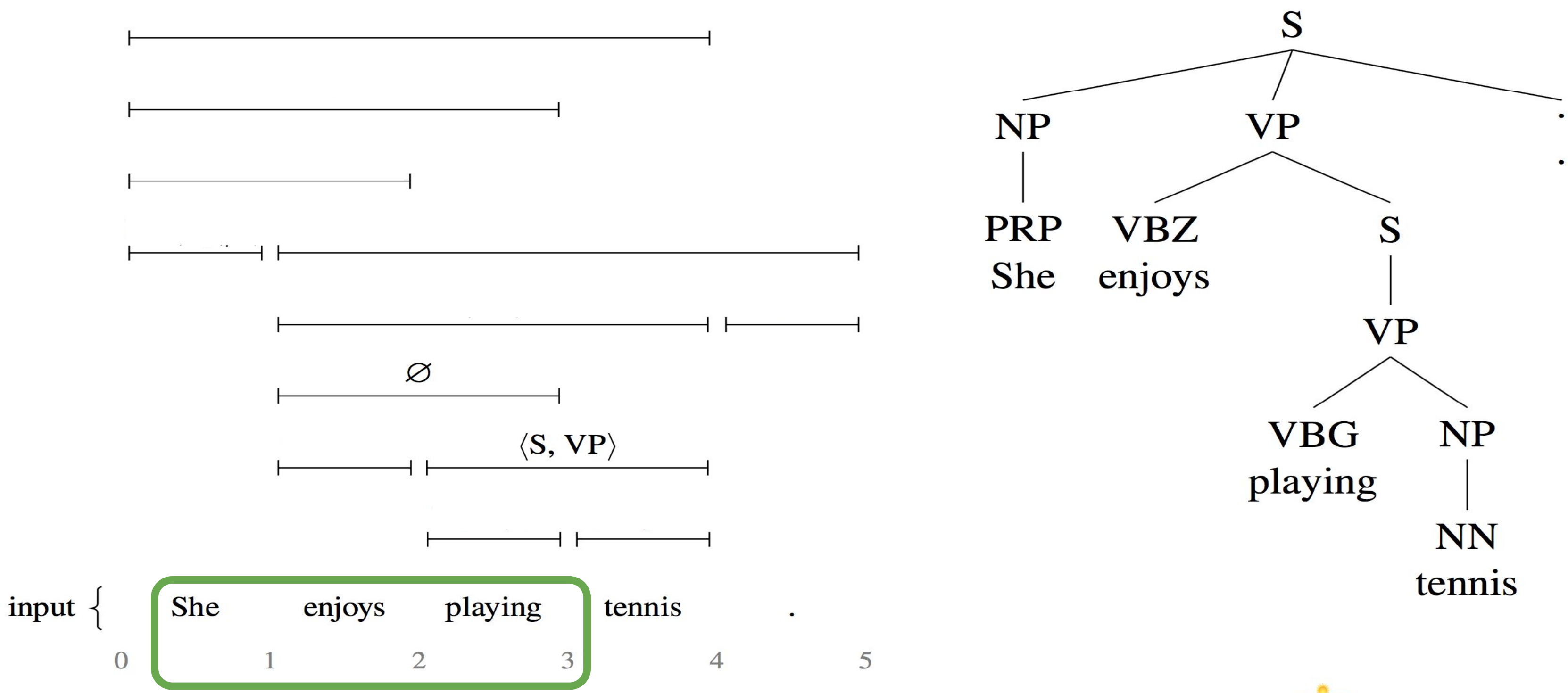
ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

Parse Tree Labels All Spans*



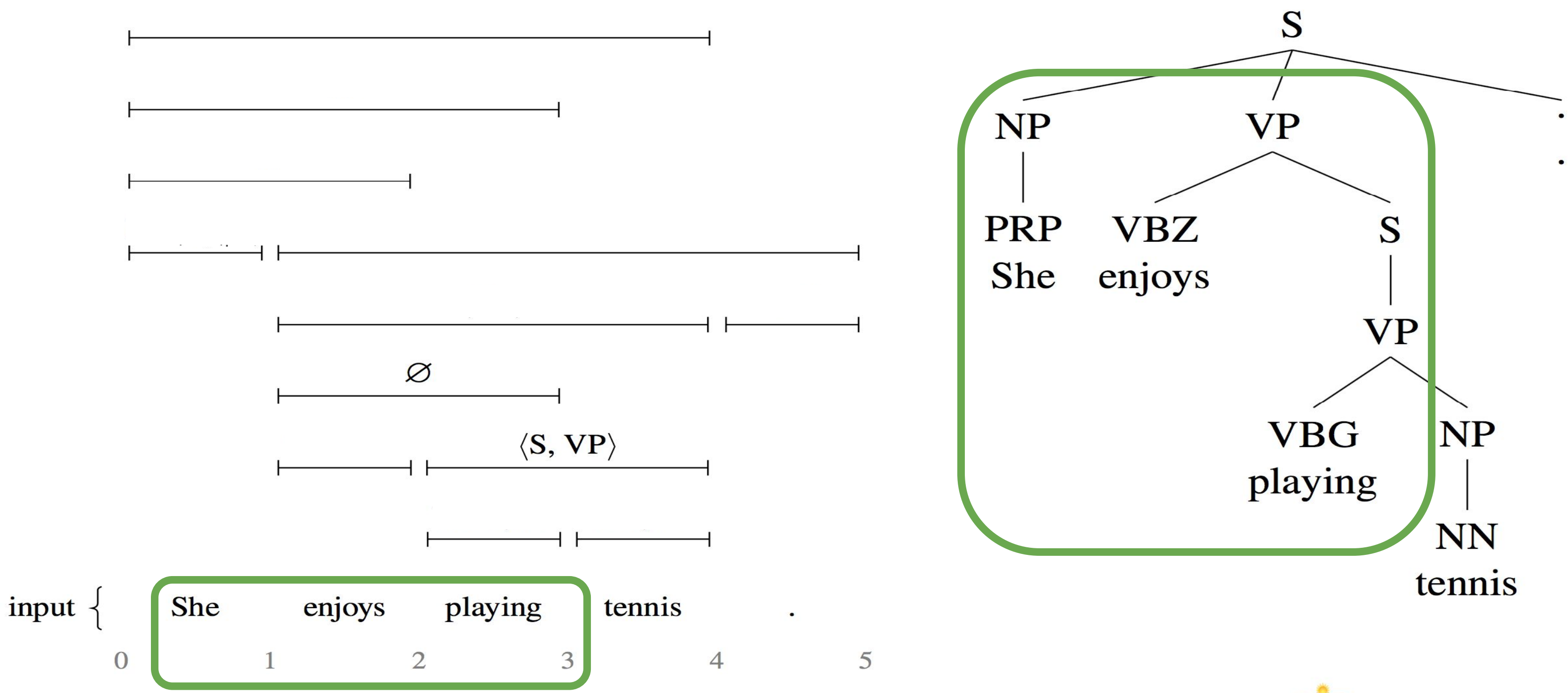
*(Cross and Huang, 2016; Stern et al., 2017)

Parse Tree Labels All Spans*



*(Cross and Huang, 2016; Stern et al., 2017)

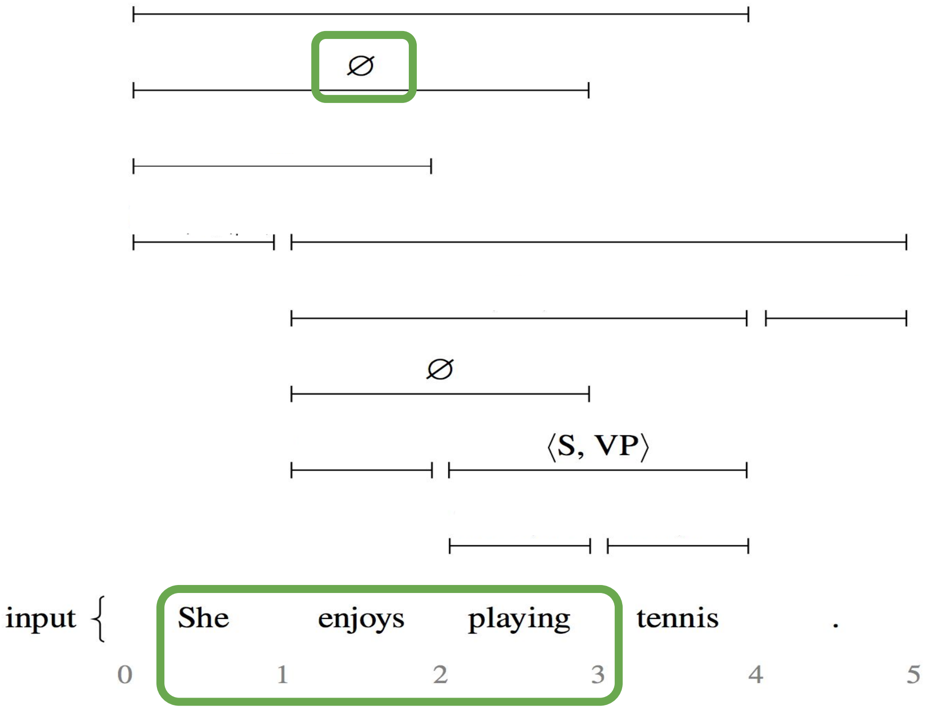
Parse Tree Labels All Spans*

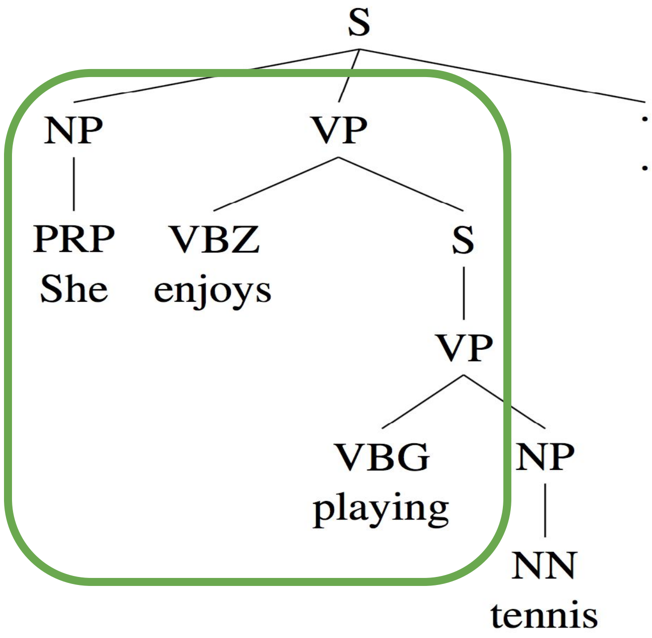


*(Cross and Huang, 2016; Stern et al., 2017)

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

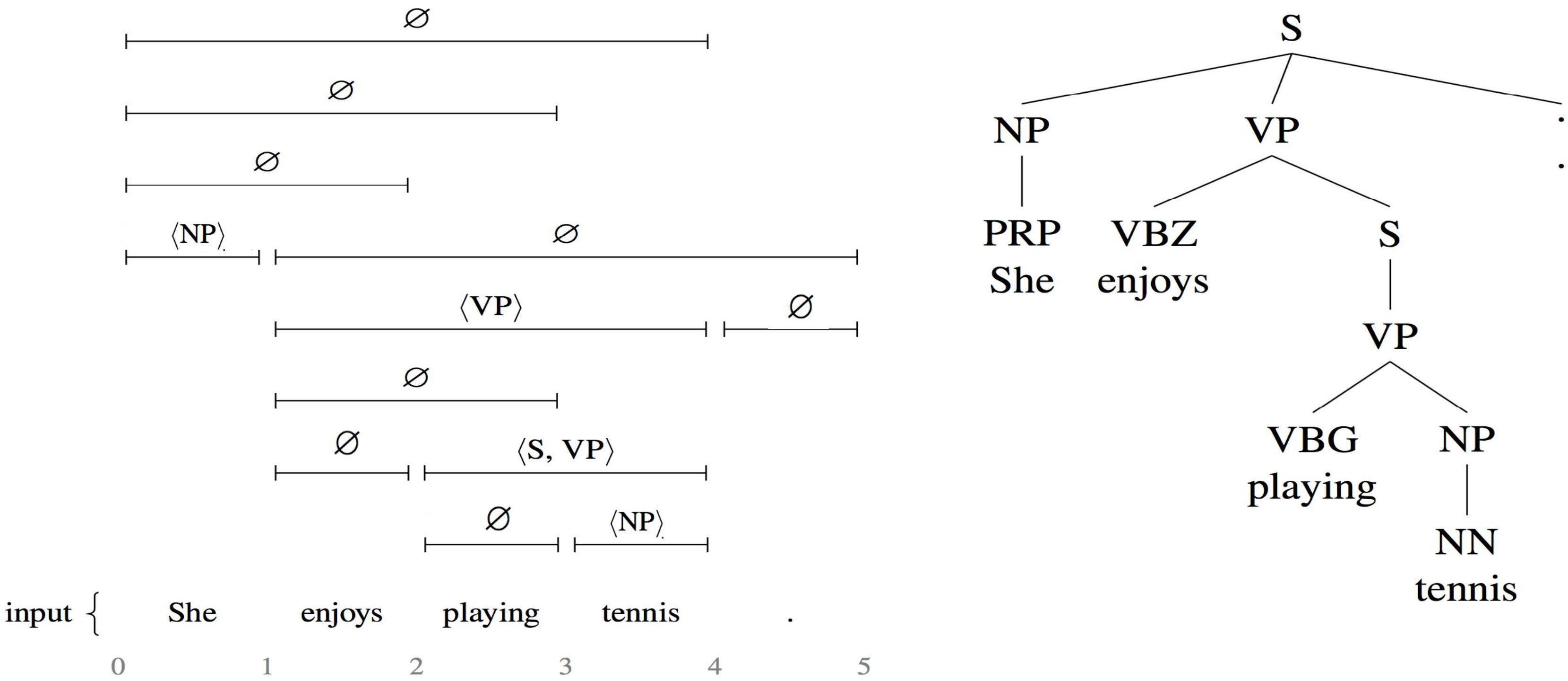
Parse Tree Labels All Spans*





*(Cross and Huang, 2016; Stern et al., 2017)

Parse Tree Labels All Spans*



*(Cross and Huang, 2016; Stern et al., 2017)

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

Training on Full and Partial Annotations

- A partial annotation is a labeled span.
- A full parse labels every span in the sentence.

Therefore, training on both is identical under our derived objective.

$$\mathcal{L}(\theta) = -\sum_{(\text{span},\text{label},\text{sentence})} \log \Pr_\theta(\text{label} \mid \text{sentence}, \text{span})$$

Parsing Using Span Classification Model

Find maximum using dynamic programming:

$$\Pr_\theta(\text{parse}|\text{sentence}) = \prod_{\text{span}\in\text{spans}} \Pr_\theta\left(\text{label of span in parse}|\text{sentence, span}\right)$$

Summary

Partial annotations are labeled spans.

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

Summary

Partial annotations are labeled spans.

Use a span classification model to parse.

Summary

Partial annotations are labeled spans.

Use a span classification model to parse.

Training on partial and full annotations becomes identical.

Definition

Training

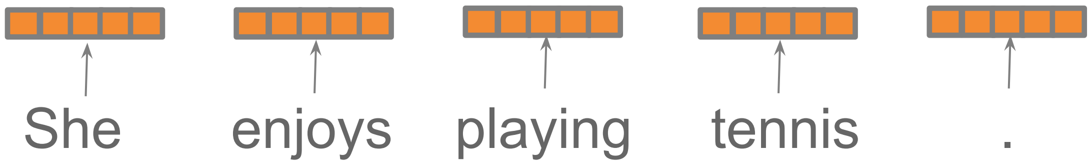
Parsing as Span Classification

The Span Classification Model

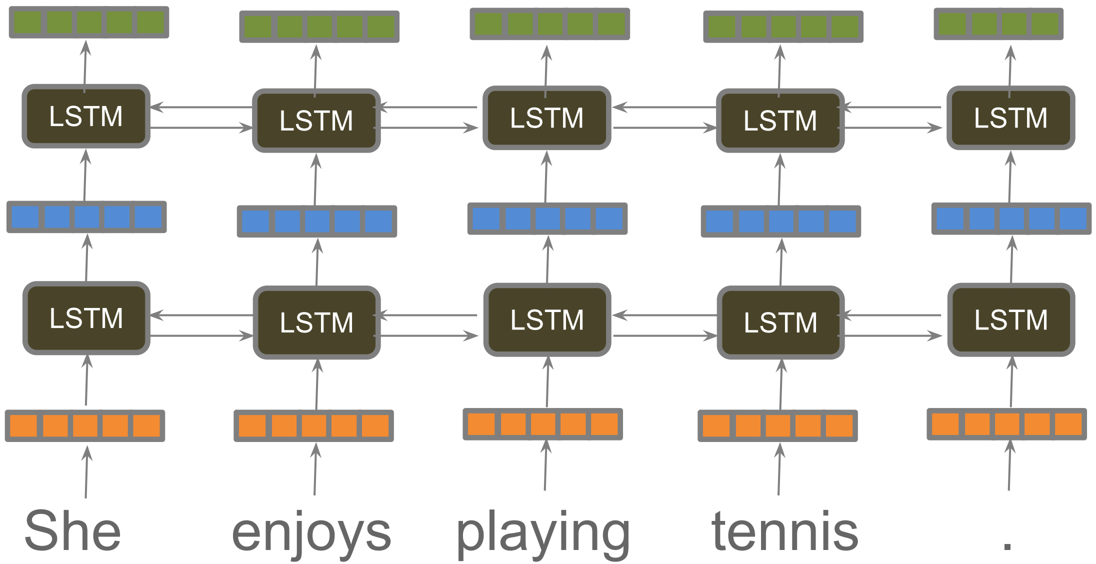
Model Architecture (Stern et al., 2017)

She enjoys playing tennis .

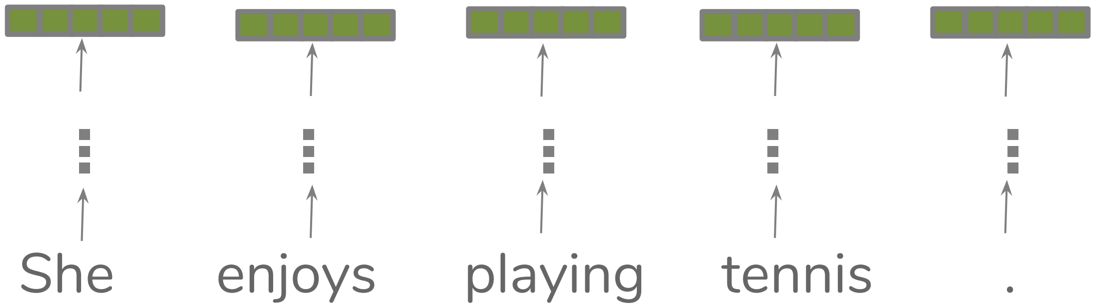
Model Architecture (Stern et al., 2017)



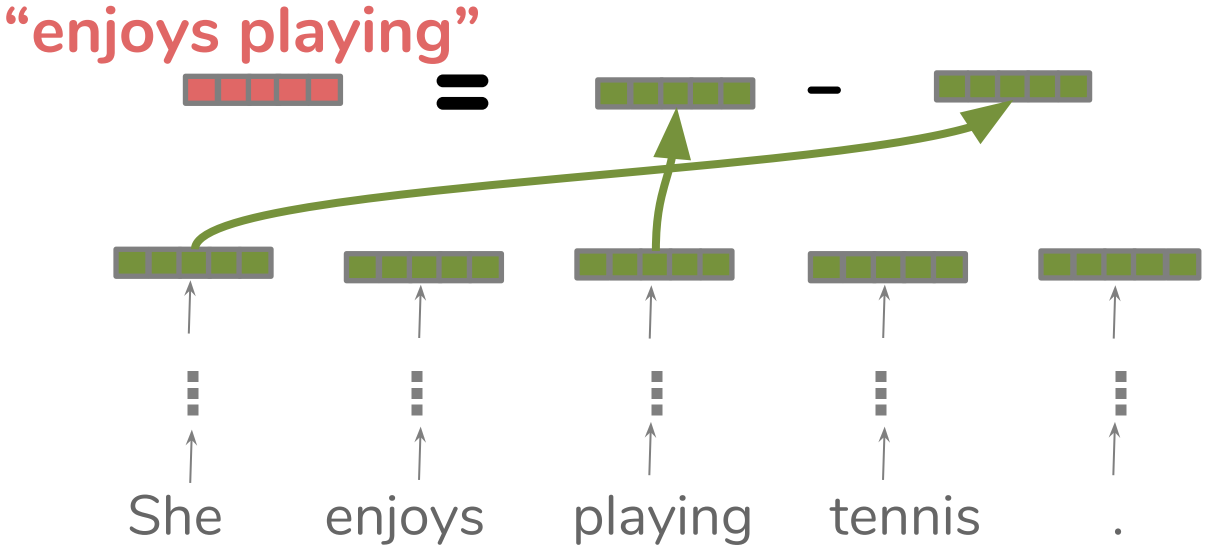
Model Architecture (Stern et al., 2017)



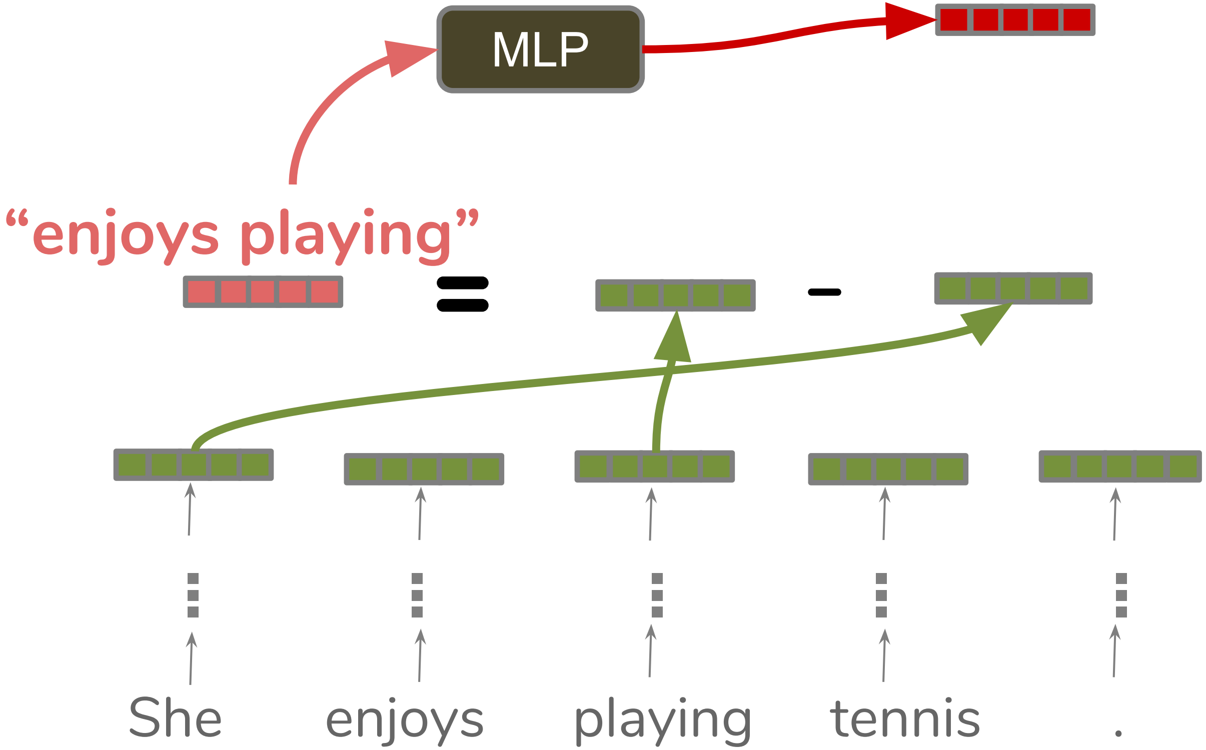
Model Architecture (Stern et al., 2017)



Span Embedding (Wang and Chang, 2016; Cross and Huang, 2016; Stern et al., 2017)



Model Architecture (Stern et al., 2017)



Differences

	Ours	Stern et al., 2017
Objective	Maximum likelihood on labels	Maximum margin on trees
ELMo	Yes	No
POS Tags as Input	No	Yes

Differences

	Ours	**Stern et al., 2017**
Objective	Maximum likelihood on labels	Maximum margin on trees
ELMo	Yes	No
POS Tags as Input	No	Yes

Differences

	Ours	**Stern et al., 2017**
Objective	Maximum likelihood on labels	Maximum margin on trees
ELMo	Yes	No
POS Tags as Input	No	Yes

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

Differences

	Ours	Stern et al., 2017
Objective	Maximum likelihood on labels	Maximum margin on trees
ELMo	Yes	No
POS Tags as Input	No	Yes

Experiments and Results

Performance on PTB

Learning Curve on New Domains

Adapting Using Partial Annotations

Performance on PTB

91.8 F1

Stern et al., 2017

+0.3 F1

+Maximum Likelihood on Labels
-POS tags

+2.2 F1

+ELMo

=

94.3 F1

Ours

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

Performance on PTB

92.6 F1

Effective Inference for Generative Neural Parsing

94.3 F1

Ours

+1.7 F1

Over Previous SoTA*

*New SoTA is 95.1 (Kitaev and Klein, ACL 2018)

Performance on PTB

Learning Curve on New Domains

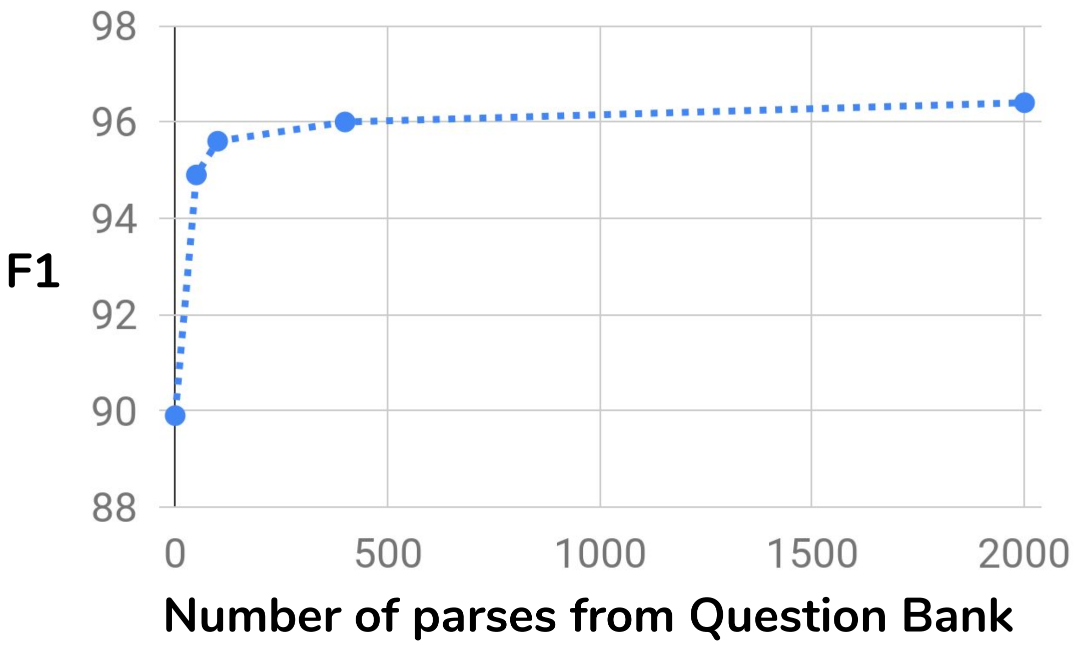
Adapting Using Partial Annotations

Question Bank (Judge et al., 2006)

- 4,000 questions.
- In contrast, PTB has few questions.

Who is the author of the book, ``The Iron Lady: A Biography of Margaret Thatcher''?

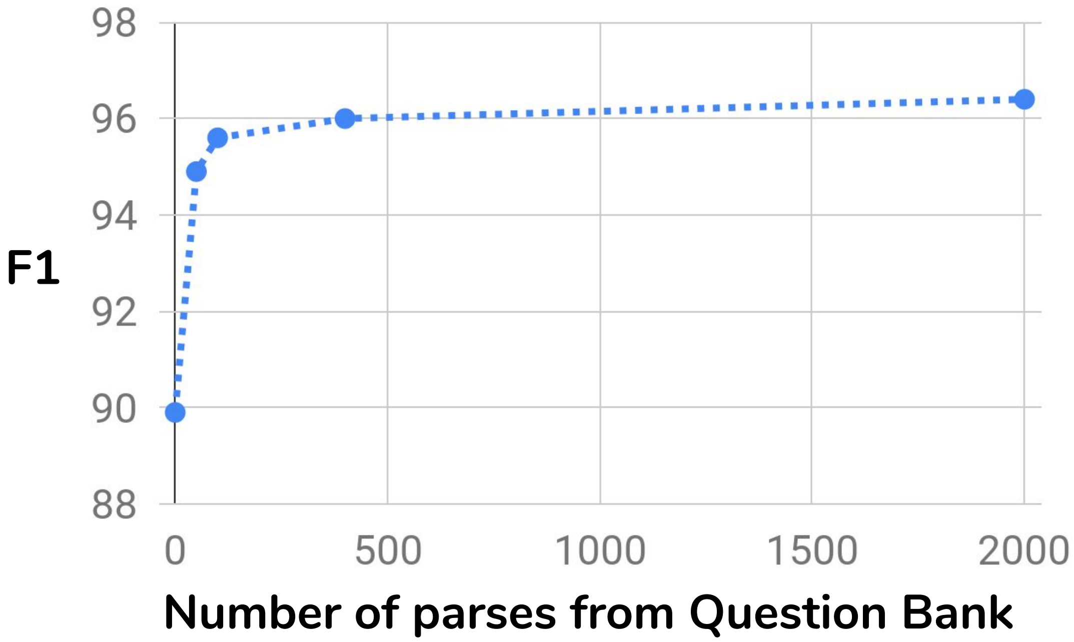
Do We Need Domain Adaptation?



+7.2 %

Training on QB

How Much Data Do We Need?



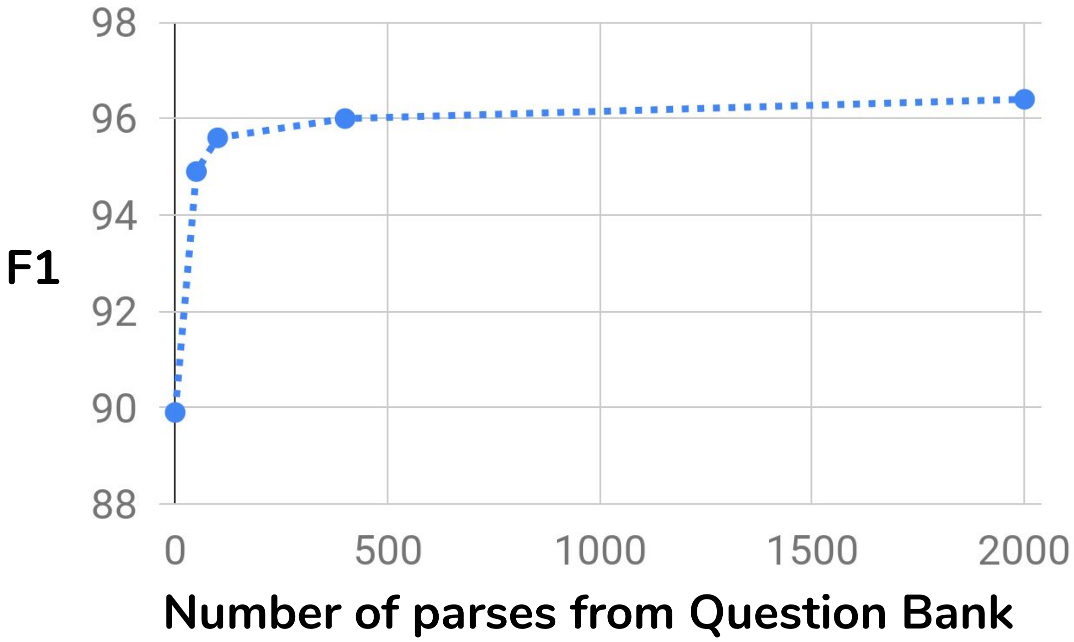
+6.3 %

From 0 to 100 parses

+0.9 %

From 100 to 2,000 parses

How Much Data Do We Need?



Not Much

Improvements taper quickly

Performance on PTB

Learning Curve on New Domains

Adapting Using Partial Annotations

Geometry Problems (Seo et al., 2015)

In the diagram at the right, circle O has a radius of 5, and CE = 2. Diameter AC is perpendicular to chord BD at E. What is the length of BD?

Biochemistry (Nivre et al., 2007)

Ethoxycoumarin was metabolized by isolated epidermal cells via dealkylation to 7-hydroxycoumarin (7-OHC) and subsequent conjugation .

Setup

Annotator is a parsing expert.

Sees parser output.

Annotated sentences randomly split into train and dev.

Biochemistry Annotations

610 partial annotations (Avg. 4.6 per sentence)

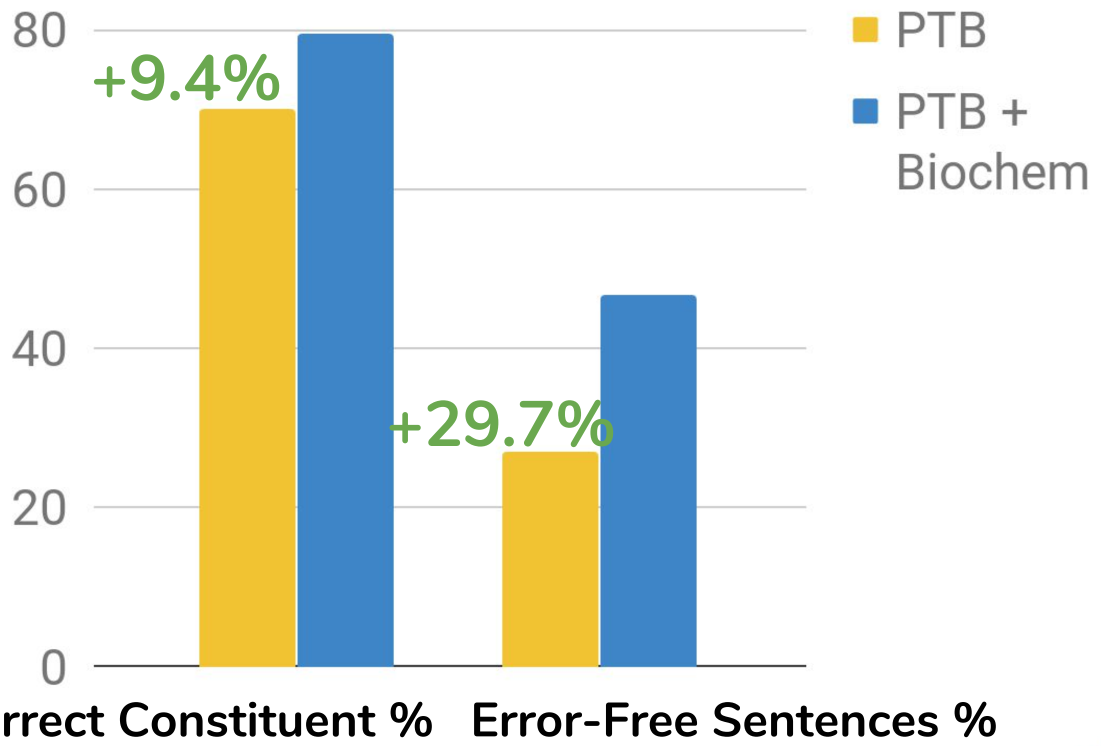
train: 72 sent, dev: 62 sent

[[In situ] hybridization] has revealed a striking subnuclear distribution of [c-myc RNA transcripts] .

[Cell growth of neuroblastoma cells in [serum containing medium]] was clearly diminished by [inhibition of FPTase]

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

What do partial annotations buy us?



ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
ALLENAI.ORG

Geometry Annotations

379 partial annotations (Avg. 3 per sentence)

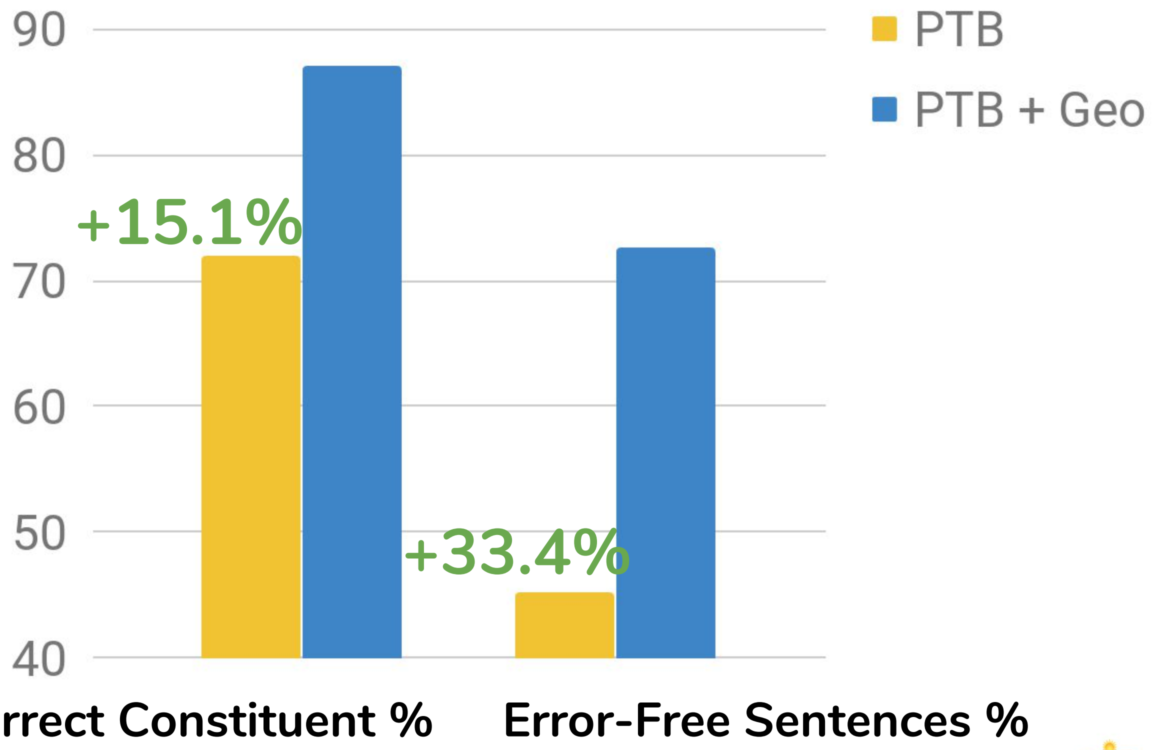
train: 63 sent, dev: 62 sent

What is [the value of [y { + z }]] ?

[Diameter AC] is perpendicular [to chord BD] [at E] .

Find [the measure of [the angle designated by x]] .

What do partial annotations buy us?



Iterative Annotation

Error Analysis on Geometry Training Set

44% math syntax

Eg: “dimensions 16 by 8,” “BAC = ¼ * ACB”

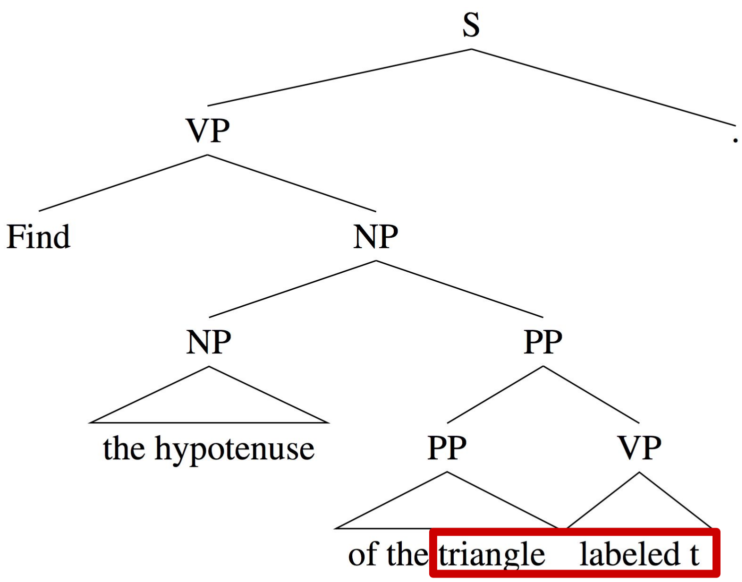
19% right-attaching participial adjectives

Eg: “segment labeled x,” “the center indicated”

19% PP-attachment

Right Attaching Participial Adjective Error

Find the hypotenuse of the triangle labeled t.



ALLENAI.ORG ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Iterative Annotation Proof-of-Concept

Invent 3 sentences similar to the incorrect one:

Find the hypotenuse of [the triangle labeled t] .

Iterative Annotation Proof-of-Concept

Invent 3 sentences similar to the incorrect one:

Find the hypotenuse of [the triangle labeled t] .

Given [a circle with [the tangent shown]] .

Iterative Annotation Proof-of-Concept

Invent 3 sentences similar to the incorrect one:

Find the hypotenuse of [the triangle labeled t] .

Given [a circle with [the tangent shown]] .

Examine [the following diagram with [the square highlighted]] .

Performance after Iterative Annotation

Correctly identified constituents:

87.0% → 88.6% (+1.6)

Error free sentences:

72.6% → 75.8% (+2.7)

Conclusion

- Recent developments make it much easier to train on partial annotations and build custom parsers.
- Making a few partial annotations can lead to significant performance improvements.

Demo: http://demo.allennlp.org/constituency-parsing

Datasets: https://github.com/vidurj/parser-adaptation/tree/master/data