Domain Adaptation for Constituency Parsing Using Partial Annotations

Vidur Joshi
Matthew Peters
Mark Hopkins
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
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.
Parse geometry sentence with PTB trained parser

\[
S \\
\downarrow \\
NP \text{ (A triangle)} \hspace{1cm} VP \hspace{1cm} NP \\
| \hspace{1cm} \downarrow \hspace{1cm} \downarrow \\
A \text{ triangle} \hspace{1cm} \text{has} \hspace{1cm} \text{NP} \\
| \hspace{1cm} \downarrow \hspace{1cm} \downarrow \\
NP \hspace{1cm} PP \\
| \hspace{1cm} \downarrow \hspace{1cm} \downarrow \\
a \text{perimeter} \hspace{1cm} \text{of 16 and one side of length 4}
\]
Parse geometry sentence with PTB trained parser
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).
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
A triangle has a perimeter of 16 and one side of length 4.
A triangle has [a perimeter of 16] and one side of length 4.
A triangle has \[\text{a perimeter of 16}\] and one side of length 4.
A triangle has \([\text{a perimeter} \{\text{of 16}\} \text{ and one side of length 4}]\).
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 \([\text{a perimeter of 16}]\) and one side of length 4.

A triangle has \([\text{NP a perimeter of 16}]\) and one side of length 4.

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

\[ L(\theta) = - \sum_{(\text{sentence}, \text{parse})} \log \Pr_{\theta}(\text{parse}|\text{sentence}) \]
Since we do not have a full parse, marginalize out components for which no supervision exists.

\[ \mathcal{L}(\theta) = - \sum_{(\text{sentence, annotations})} \log \left( \sum_{\text{parses consistent with annotations}} \Pr_{\theta}(\text{parse} | \text{sentence}) \right) \]
Objective for Partial Annotation

Marginalize out components for which no supervision exists.

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

Expensive!
One Solution: Approximation*

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

*(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.

\[
Pr_\theta (\text{parse}|\text{sentence}) = \prod_{(\text{span}, \text{label}) \text{ consistent with parse}} Pr_\theta (\text{label}|\text{sentence, span})
\]
Our Solution: Parsing as Span Classification

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

\[
\Pr_\theta(\text{parse}|\text{sentence}) = \prod_{(\text{span},\text{label}) \text{ consistent with parse}} \Pr_\theta(\text{label}|\text{sentence}, \text{span})
\]
Our Solution: Parsing as Span Classification

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

\[ \text{Pr}_\theta(\text{parse}|\text{sentence}) = \prod_{(\text{span, label}) \text{ consistent with parse}} \text{Pr}_\theta(\text{label}|\text{sentence, span}) \]

Objective now simplifies to:

\[ \mathcal{L}(\theta) = - \sum_{(\text{sentence, annotations})} \sum_{(\text{span, label}) \in \text{annotations}} \log \text{Pr}_\theta(\text{label}|\text{sentence, span}) \]

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)*
*(Cross and Huang, 2016; Stern et al., 2017)
Parse Tree Labels All Spans*

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

\[ L(\theta) = - \sum_{(\text{span}, \text{label}, \text{sentence})} \log \Pr_{\theta}(\text{label}|\text{sentence}, \text{span}) \]
Parsing Using Span Classification Model

Find maximum using dynamic programming:

$$\Pr_\theta(\text{parse} | \text{sentence}) = \prod_{\text{span} \in \text{spans}} \Pr_\theta(\text{label of span in parse} | \text{sentence, span})$$
Partial annotations are labeled spans.
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.
She enjoys playing tennis.
Model Architecture (Stern et al., 2017)

She enjoys playing tennis.
Model Architecture (Stern et al., 2017)

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

“enjoys playing”

She enjoys playing tennis.
She enjoys playing tennis.
## Differences

<table>
<thead>
<tr>
<th></th>
<th>Ours</th>
<th>Stern et al., 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objective</strong></td>
<td>Maximum likelihood on labels</td>
<td>Maximum margin on trees</td>
</tr>
<tr>
<td><strong>ELMo</strong></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>POS Tags as Input</strong></td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
### Differences

<table>
<thead>
<tr>
<th></th>
<th>Ours</th>
<th>Stern et al., 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objective</strong></td>
<td>Maximum likelihood on labels</td>
<td>Maximum margin on trees</td>
</tr>
<tr>
<td><strong>ELMo</strong></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>POS Tags as Input</strong></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>Stern et al., 2017</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------</td>
<td>--------------------</td>
</tr>
<tr>
<td><strong>Objective</strong></td>
<td>Maximum likelihood on labels</td>
<td>Maximum margin on trees</td>
</tr>
<tr>
<td><strong>ELMo</strong></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>POS Tags as Input</strong></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>Stern et al., 2017</td>
</tr>
<tr>
<td>------------------------</td>
<td>-----------------------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td><strong>Objective</strong></td>
<td>Maximum likelihood on labels</td>
<td>Maximum margin on trees</td>
</tr>
<tr>
<td><strong>ELMo</strong></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>POS Tags as Input</strong></td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
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

= 94.3 F1

+2.2 F1
+ELMo

Ours
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
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?

![Graph showing F1 score improvement with increasing number of parses from Question Bank]

F1

+7.2 %

Training on QB

Number of parses from Question Bank

88 90 92 94 96 98

0 500 1000 1500 2000
How Much Data Do We Need?

![Graph showing F1 score improvement with number of parses.](image)

- **+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
Experiments and Results

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.
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.
What do partial annotations buy us?

- Correct Constituent %: +9.4%
- Error-Free Sentences %: +29.7%
What is \( [ \text{the value of (} y \{ + z \} ) ] \)?

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

Find [ the measure of [ the angle designated by x ] ].
What do partial annotations buy us?

- Correct Constituent %: +15.1%
- Error-Free Sentences %: +33.4%
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
Find the hypotenuse of the triangle labeled t.
Invent 3 sentences similar to the incorrect one:

Find the hypotenuse of the triangle labeled t.
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