Incremental Syntactic Language Models for Phrase-based Translation

Lane Schwartz
Chris Callison-Burch  William Schuler  Stephen Wu

Air Force Research Lab
lane.schwartz@wpafb.af.mil

Johns Hopkins University
ccb@cs.jhu.edu

Ohio State University
schuler@ling.ohio-state.edu

Mayo Clinic
wu.stephen@mayo.edu
Translation Model vs Language Model
Syntax in the Translation Model

Abeillé et al., 1990; Poutsma, 1998; Poutsma, 2000; Yamada & Knight, 2001; Yamada & Knight, 2002; Eisner, 2003; Gildea, 2003; Hearne & Way, 2003; Poutsma, 2003; Imamura et al., 2004; Galley et al., 2004; Graehl & Knight, 2004; Melamed, 2004; Ding & Palmer, 2005; Hearne, 2005; Quirk et al., 2005; Cowan et al., 2006; Galley et al., 2006; Huang et al., 2006; Liu et al., 2006; Marcu et al., 2006; Zollmann & Venugopal, 2006; Bod, 2007; DeNeefe et al., 2007; Liu et al., 2007; Chiang et al., 2008; Lavie et al., 2008; Mi & Huang, 2008; Mi et al., 2008; Resnik, 2008; Shen et al., 2008; Zhou et al., 2008; Chiang, 2009; Hanneman & Lavie, 2009; Liu et al., 2009; Chiang, 2010; Huang & Mi, 2010; . . .
Translation Model vs **Language Model**
An **incremental syntactic language model** uses an incremental statistical parser to define a probability model over the dependency or phrase structure of target language strings.
An **incremental syntactic language model** uses an incremental statistical parser to define a probability model over the dependency or phrase structure of target language strings.

- Phrase-based decoder produces translation in the target language incrementally from left-to-right
- Phrase-based syntactic LM parser should parse target language hypotheses incrementally from left-to-right
- Related work:
  - Galley & Manning (2009) obtained 1-best dependency parse using a greedy dependency parser
An **incremental syntactic language model** uses an incremental statistical parser to define a probability model over the dependency or phrase structure of target language strings.

- Phrase-based decoder produces translation in the target language incrementally from left-to-right
- Phrase-based syntactic LM parser should parse target language hypotheses incrementally from left-to-right

**Related work:**
- Galley & Manning (2009) obtained 1-best dependency parse using a greedy dependency parser
An incremental syntactic language model uses an incremental statistical parser to define a probability model over the dependency or phrase structure of target language strings.

- Phrase-based decoder produces translation in the target language incrementally from left-to-right
- Phrase-based syntactic LM parser should parse target language hypotheses incrementally from left-to-right
- Related work:
  - Galley & Manning (2009) obtained 1-best dependency parse using a greedy dependency parser
An incremental syntactic language model uses an incremental statistical parser to define a probability model over the dependency or phrase structure of target language strings.

We use a standard HHMM parser (Schuler et al., 2010)

**Engineering** simple model, equivalent to PPDA

**Engineering** linear-time parsing

**Algorithmic** elegant fit into phrase-based decoder

**Cognitive** nice psycholinguistic properties

**Other parsers** Roark (2001), Henderson (2004), Huang & Sagae (2010)
An incremental syntactic language model uses an incremental statistical parser to define a probability model over the dependency or phrase structure of target language strings.

We use a standard HHMM parser (Schuler et al., 2010)

**Engineering** simple model, equivalent to PPDA

**Engineering** linear-time parsing

**Algorithmic** elegant fit into phrase-based decoder

**Cognitive** nice psycholinguistic properties

**Other parsers** Roark (2001), Henderson (2004), Huang & Sagae (2010)
The president meets the board on Friday.
Incremental Parsing

Motivation

Syntactic LM

Decoder Integration

Results

Questions?

Incremental Syntactic Language Models for Phrase-based Translation

Lane Schwartz
Transform right-expanding sequences of constituents into left-expanding sequences of incomplete constituents (Johnson 1998)

Incomplete constituents can be processed incrementally using a Hierarchical Hidden Markov Model parser. (Murphy & Paskin, 2001; Schuler et al. 2010)
Transform right-expanding sequences of constituents into left-expanding sequences of incomplete constituents (Johnson 1998)

Incomplete constituents can be processed incrementally using a Hierarchical Hidden Markov Model parser. (Murphy & Paskin, 2001; Schuler et al. 2010)
Transform right-expanding sequences of constituents into left-expanding sequences of incomplete constituents (Johnson 1998)

Incomplete constituents can be processed incrementally using a Hierarchical Hidden Markov Model parser. (Murphy & Paskin, 2001; Schuler et al. 2010)
Incremental Parsing using HHMM (Schuler et al. 2010)

Hierarchical Hidden Markov Model

- Circles denote hidden random variables
- Edges denote conditional dependencies
- Shaded circles denote observed values

Motivation

Syntactic LM

Decoder Integration

Results

Questions?

Incremental Syntactic Language Models for Phrase-based Translation

Lane Schwartz
Incremental Parsing using HHMM (Schuler et al. 2010)

Hierarchical Hidden Markov Model

- Circles denote hidden random variables
- Edges denote conditional dependencies
- Shaded circles denote observed values

Motivation
Syntactic LM
Decoder Integration
Results
Questions?

Incremental Syntactic Language Models for Phrase-based Translation

Lane Schwartz
Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path

Motivation

Syntactic LM

Decoder Integration

Results

Questions?

Incremental Syntactic Language Models for Phrase-based Translation

Lane Schwartz
Incremental Parsing using HHMM (Schuler et al. 2010)

**Motivation**

**Syntactic LM**

**Decoder Integration**

**Results**

**Questions?**

Incremental Syntactic Language Models for Phrase-based Translation

Lane Schwartz

---

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path

---

```
S
/ \   /
S/VP S/NP NP
|    |   /
VP  IN \ on
/   /
|  |
NP/NN VP/NN NN
|  |
DT VP/NP DT
president the board
|  |
NN VP/NP DT
meets the board
|  |
IN NN VP/NP
on the board
|  |
IN NP/NN VP/NP
on the board

The president meets the board on Friday
```

---

```
s_1 = The
s_2 = president
s_3 = meets
s_4 = the
s_5 = board
s_6 = on
s_7 = Friday
```

---

```
r_1
/ |
r_2
/ |
r_3
/ |
r_4
/ |
r_5
/ |
r_6
/ |
r_7
/ |
r_8
/ |
```

---

```
e_1 = The
e_2 = president
e_3 = meets
e_4 = the
e_5 = board
e_6 = on
e_7 = Friday
```

---

```
r_1
|   |
r_2
|   |
r_3
|   |
r_4
|   |
r_5
|   |
r_6
|   |
r_7
|   |
r_8
|   |
```

---

```
s_1
/ |
s_2
/ |
s_3
/ |
s_4
/ |
s_5
/ |
s_6
/ |
s_7
/ |
s_8
/ |
```

---

```
r_1
|   |
r_2
|   |
r_3
|   |
r_4
|   |
r_5
|   |
r_6
|   |
r_7
|   |
r_8
|   |
```

---

```
e_1 = The
e_2 = president
e_3 = meets
e_4 = the
e_5 = board
e_6 = on
e_7 = Friday
```

---

```
s_1
/ |
s_2
/ |
s_3
/ |
s_4
/ |
s_5
/ |
s_6
/ |
s_7
/ |
s_8
/ |
```

---

```
r_1
|   |
r_2
|   |
r_3
|   |
r_4
|   |
r_5
|   |
r_6
|   |
r_7
|   |
r_8
|   |
```

---

```
e_1 = The
e_2 = president
e_3 = meets
e_4 = the
e_5 = board
e_6 = on
e_7 = Friday
```
Incremental Parsing using HHMM (Schuler et al. 2010)

Analogous to “Maximally Incremental”
CCG Parsing

Equivalent to Probabilistic
Push-Down Automata

Isomorphic Tree → Path

Motivation

Syntactic LM

Decoder Integration

Results

Questions?

Incremental Syntactic Language Models for Phrase-based Translation

Lane Schwartz
Incremental Parsing using HHMM (Schuler et al. 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path
Incremental Parsing using HHMM (Schuler et al. 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path
Incremental Parsing using HHMM (Schuler et al. 2010)

Analogous to “Maximally Incremental”
CCG Parsing

Equivalent to Probabilistic
Push-Down Automata

Isomorphic Tree → Path
Analogous to “Maximally Incremental”
CCG Parsing

Equivalent to Probabilistic
Push-Down Automata

Isomorphic Tree → Path
Incremental Parsing using HHMM (Schuler et al. 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path

Motivation

Syntactic LM

Decoder Integration

Results

Questions?

Incremental Syntactic Language Models for Phrase-based Translation

Lane Schwartz
Incremental Parsing using HHMM (Schuler et al. 2010)

Analogous to “Maximally Incremental”
CCG Parsing

Equivalent to Probabilistic
Push-Down Automata

Isomorphic Tree → Path
Analogous to “Maximally Incremental” CCG Parsing

Isomorphisc Tree → Path
Incremental Parsing using HHMM (Schuler et al. 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path
Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path
Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path
Analogous to “Maximally Incremental”
CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path

Incremental Syntactic Language Models for Phrase-based Translation
Lane Schwartz
Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path
Incremental Parsing using HHMM (Schuler et al. 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path
Incremental Parsing using HHMM (Schuler et al. 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path
Phrase-Based Translation

Der Präsident trifft am Freitag den Vorstand
The president meets the board on Friday
Der Präsident trifft am Freitag den Vorstand
The president meets the board on Friday
Definition

\[ \tilde{\tau}_{th} \] represents parses of the partial translation at node \( h \) in stack \( t \)
Definition

\( \tilde{\tau}_{th} \) represents parses of the partial translation at node \( h \) in stack \( t \)
Integrate Parser into Phrase-based Decoder

Motivation

Syntactic LM

Decoder Integration

Results

Questions?

Incremental Syntactic Language Models for Phrase-based Translation

Lane Schwartz
Integrate Parser into Phrase-based Decoder

Motivation
Syntactic LM
Decoder Integration
Results
Questions?

Incremental Syntactic Language Models for Phrase-based Translation

Lane Schwartz
Direct Maximum Entropy Model of Translation

\[ \hat{e} = \arg\max_e \exp \sum_j \lambda_j h_j(e, f) \]

\( \lambda \) = Set of \( j \) feature weights

\( h \) = \{ Phrase-based translation model, \n-gram LM, Distortion model \, ... \}

Syntactic LM \( P(\tilde{\tau}_{th}) \)

Stack 0

\[ \langle s \rangle \tilde{\tau}_0 \]

Stack 1

\[ \langle s \rangle \text{the} \tilde{\tau}_{1_1} \]

Stack 2

\[ \text{the president} \tilde{\tau}_{2_1} \]

Stack 3

\[ \text{president meets} \tilde{\tau}_{3_1} \]
That’s nice...

but will it make my BLEU score go up?
Perplexity Results

Language models trained on WSJ Treebank corpus

<table>
<thead>
<tr>
<th>LM</th>
<th>In-domain Perplexity</th>
<th>Out-of-domain Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ 5-gram LM</td>
<td>232</td>
<td>1262</td>
</tr>
<tr>
<td>WSJ Syntactic LM</td>
<td>385</td>
<td>529</td>
</tr>
</tbody>
</table>

Motivation

Incremental Syntactic Language Models for Phrase-based Translation

Syntactic LM

Decoder Integration

Results

Lane Schwartz
### Perplexity Results

Language models trained on WSJ Treebank corpus

<table>
<thead>
<tr>
<th>LM</th>
<th>In-domain Perplexity</th>
<th>Out-of-domain Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ 5-gram LM</td>
<td>232</td>
<td>1262</td>
</tr>
<tr>
<td>WSJ Syntactic LM</td>
<td>385</td>
<td>529</td>
</tr>
<tr>
<td>Interpolated</td>
<td>209</td>
<td>225</td>
</tr>
<tr>
<td>WSJ 5-gram + WSJ SynLM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Language models trained on WSJ Treebank corpus
...and \( n \)-gram model for larger English Gigaword corpus.

<table>
<thead>
<tr>
<th>LM</th>
<th>In-domain Perplexity</th>
<th>Out-of-domain Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSJ 5-gram LM</td>
<td>232</td>
<td>1262</td>
</tr>
<tr>
<td>WSJ Syntactic LM</td>
<td>385</td>
<td>529</td>
</tr>
<tr>
<td>Interpolated</td>
<td>209</td>
<td>225</td>
</tr>
<tr>
<td>WSJ 5-gram + WSJ SynLM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gigaword 5-gram</td>
<td>258</td>
<td>312</td>
</tr>
<tr>
<td>Interpolated</td>
<td>222</td>
<td>123</td>
</tr>
<tr>
<td>Gigaword 5-gram + WSJ SynLM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
That’s nice...

but will it make my BLEU score go up?
Does an Incremental Syntactic LM Help Translation?

<table>
<thead>
<tr>
<th>Moses with LM(s)</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using $n$-gram LM only</td>
<td>18.78</td>
</tr>
<tr>
<td>Using $n$-gram LM + Syntactic LM</td>
<td>19.78</td>
</tr>
</tbody>
</table>

**Experiment**

- NIST OpenMT 2008 Urdu-English data set
- Moses with standard phrase-based translation model
- Tuning and testing restricted to sentences $\leq$ 20 words long
- Results reported on devtest set
- $n$-gram LM is WSJ 5-gram LM
Does an Incremental Syntactic LM Help Translation?

<table>
<thead>
<tr>
<th>Moses with LM(s)</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using (n)-gram LM only</td>
<td>18.78</td>
</tr>
<tr>
<td>Using (n)-gram LM + Syntactic LM</td>
<td>19.78</td>
</tr>
</tbody>
</table>

**Experiment**

- NIST OpenMT 2008 Urdu-English data set
- Moses with standard phrase-based translation model
- Tuning and testing restricted to sentences \(\leq 20\) words long
- Results reported on devtest set
- \(n\)-gram LM is WSJ 5-gram LM
Summary

- Straightforward general framework for incorporating any Incremental Syntactic LM into Phrase-based Translation
- We used an Incremental HHMM Parser as Syntactic LM
  - Syntactic LM shows substantial decrease in perplexity on out-of-domain data over \( n \)-gram LM when trained on same data
  - Syntactic LM interpolated with \( n \)-gram LM shows even greater decrease in perplexity on both in-domain and out-of-domain data, even when \( n \)-gram LM is trained on substantially larger corpus
  - +1 BLEU on Urdu-English task with Syntactic LM
- All code is open source and integrated into Moses
Incremental Syntactic Language Models for Phrase-based Translation

Lane Schwartz
Chris Callison-Burch  William Schuler  Stephen Wu

Air Force Research Lab
lane.schwartz@wpafb.af.mil

Johns Hopkins University
ccb@cs.jhu.edu

Ohio State University
schuler@ling.ohio-state.edu

Mayo Clinic
wu.stephen@mayo.edu

Motivation  Syntactic LM  Decoder Integration  Results  Questions?

Incremental Syntactic Language Models for Phrase-based Translation

Lane Schwartz
This looks a lot like CCG

Our parser performs some CCG-style operations:

- **Forward function application**
  - NP/NN NN ⇒ NP

- **Type raising**
  - NP ⇒ S/VP

- **Type raising in conjunction with forward function composition**
  - DT ⇒ NP/NN
  - VP/NP NP/NN ⇒ VP/NN
Why not just use CCG?

- No probabilistic version of incremental CCG
- Our parser is constrained (we don’t have backward composition)
- We do use those components of CCG (forward function application and forward function composition) which are useful for probabilistic incremental parsing
Speed Results

Mean per-sentence decoding time

<table>
<thead>
<tr>
<th>Sentence length</th>
<th>Moses</th>
<th>+SynLM beam=50</th>
<th>+SynLM beam=2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.2 sec</td>
<td>9 min</td>
<td>19 min</td>
</tr>
<tr>
<td>20</td>
<td>0.5 sec</td>
<td>20 min</td>
<td>43 min</td>
</tr>
<tr>
<td>30</td>
<td>0.9 sec</td>
<td>29 min</td>
<td>62 min</td>
</tr>
<tr>
<td>40</td>
<td>1.1 sec</td>
<td>35 min</td>
<td>76 min</td>
</tr>
</tbody>
</table>

- Parser beam sizes are indicated for the syntactic LM
- Parser runs in linear time, but we’re parsing all paths through the Moses lattice as they are generated by the decoder
- More informed pruning, but slower decoding
**Definition**

- \( e \) \( \overset{\text{def}}{=} \) string of \( n \) target language words \( e_1 \ldots e_n \)
- \( e_t \) \( \overset{\text{def}}{=} \) the first \( t \) words in \( e \), where \( t \leq n \)
- \( \tau_t \) \( \overset{\text{def}}{=} \) set of all incremental parses of \( e_t \)
- \( \tilde{\tau}_t \) \( \overset{\text{def}}{=} \) subset of parses \( \tau_t \) that remain after parser pruning

\[
\begin{align*}
\argmax_{\tau} P(\tau | e) & \rightarrow \hat{\tau} \\
\tilde{\tau}_{t-1} & \xrightarrow{\delta} \tilde{\tau}_t \\
\end{align*}
\]
Acknowledgments

This research was supported by NSF CAREER/PECASE award 0447685, NSF grant IIS-0713448, and the European Commission through the EuroMatrixPlus project.

Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the sponsors or the United States Air Force.

Paper cleared for public release (Case Number 88ABW-2010-6489) on 10 Dec 2010.

Presentation cleared for public release (Case Number 88ABW-2011-2970) on 26 May 2011.