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

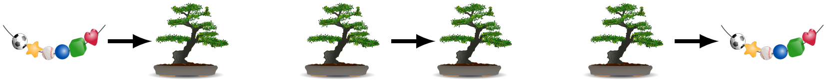
Ohio State University
schuler@ling.ohio-state.edu

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

Syntax in the Language Model

Definition

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.

Syntax in the Language Model

Definition

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

Definition

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

Syntax in the Language Model

Definition

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

Syntax in the Language Model

Definition

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)

Syntax in the Language Model

Definition

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

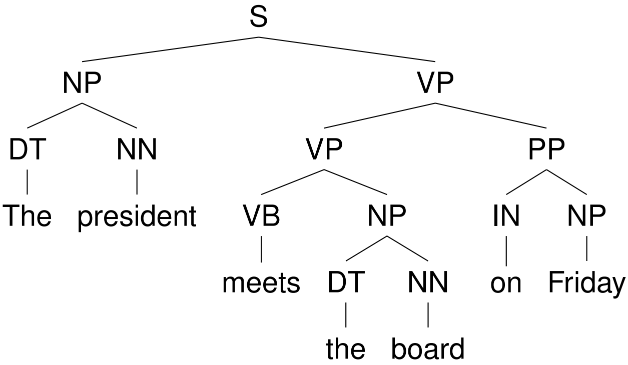
Engineering linear-time parsing

Algorithmic elegant fit into phrase-based decoder

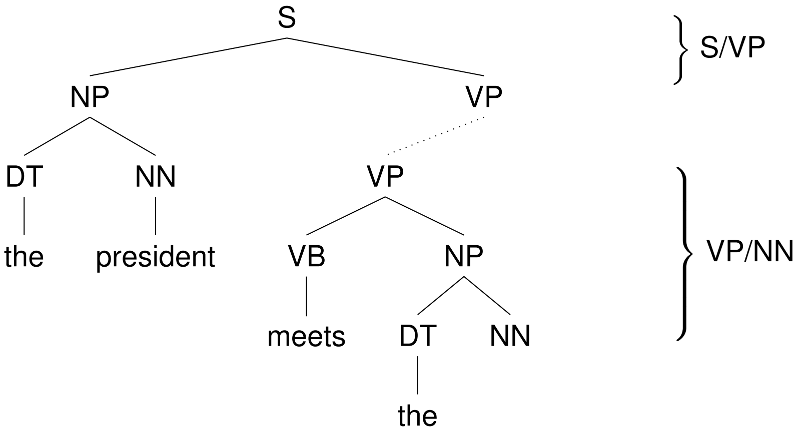
Cognitive nice psycholinguistic properties

Other parsers Roark (2001), Henderson (2004), Huang & Sagae (2010)

Incremental Parsing

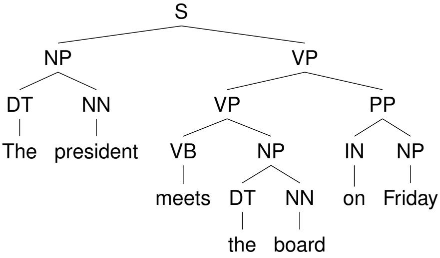


Incremental Parsing



Incremental Parsing using HHMM (Schuler *et al.* 2010)

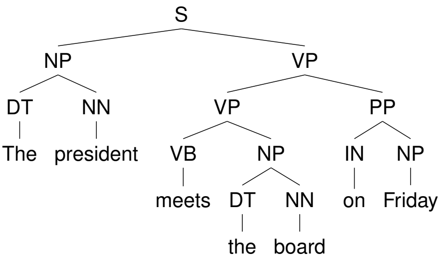
Transform right-expanding sequences of constituents into left-expanding sequences of incomplete constituents (Johnson 1998)

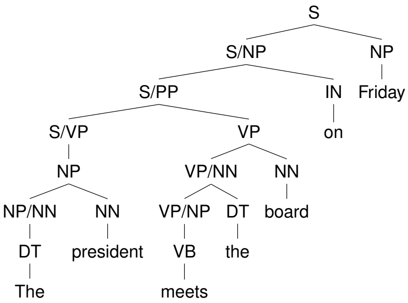


Incomplete constituents can be processed incrementally using a Hierarchical Hidden Markov Model parser. (Murphy & Paskin, 2001; Schuler *et al.* 2010)

Incremental Parsing using HHMM (Schuler *et al.* 2010)

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

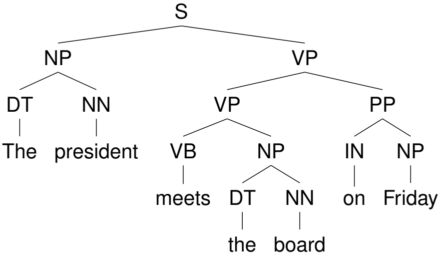


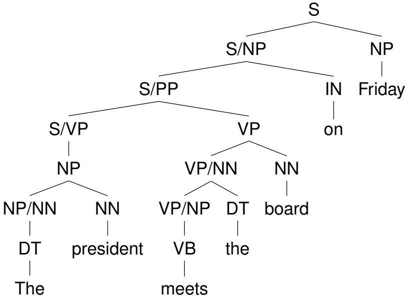


Incomplete constituents can be processed incrementally using a Hierarchical Hidden Markov Model parser. (Murphy & Paskin, 2001; Schuler *et al.* 2010)

Incremental Parsing using HHMM (Schuler *et al.* 2010)

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



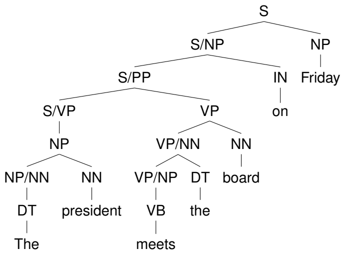


Incomplete constituents can be processed incrementally using a Hierarchical Hidden Markov Model parser. (Murphy & Paskin, 2001; Schuler *et al.* 2010)

Incremental Parsing using HHMM (Schuler *et al.* 2010)

Hierarchical Hidden Markov Model

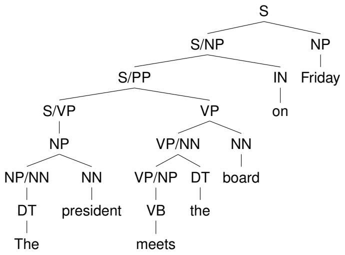
- Circles denote
  hidden random variables
- Edges denote
  conditional dependencies
- Shaded circles denote
  observed values

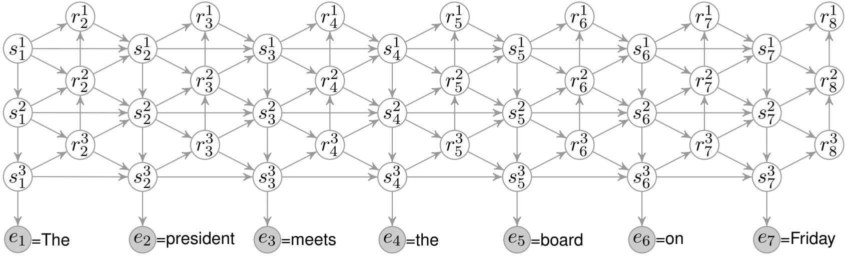


Incremental Parsing using HHMM (Schuler *et al.* 2010)

Hierarchical Hidden Markov Model

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



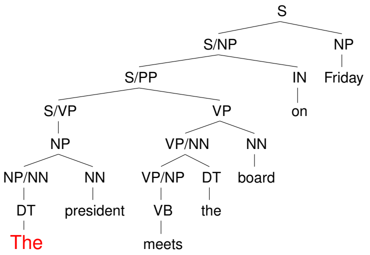


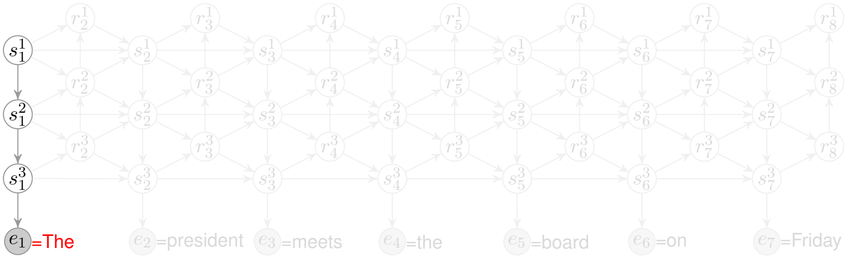
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to "Maximally Incremental" CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path

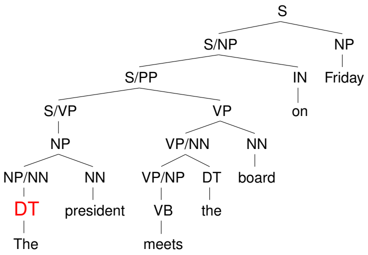


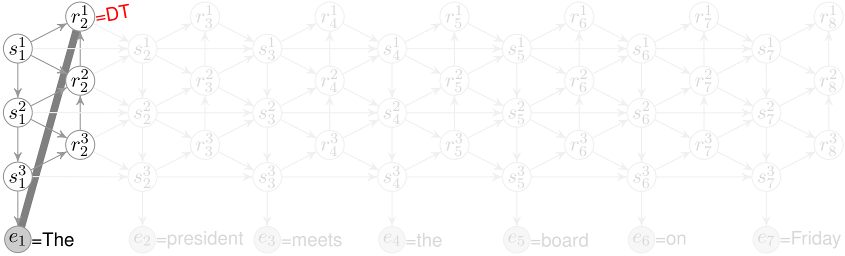


Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree $\rightarrow$ Path



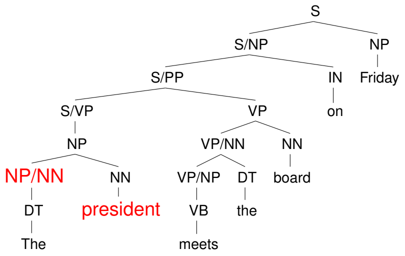


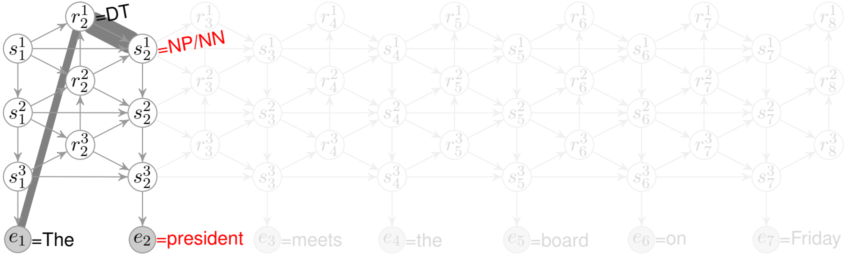
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to "Maximally Incremental" CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path



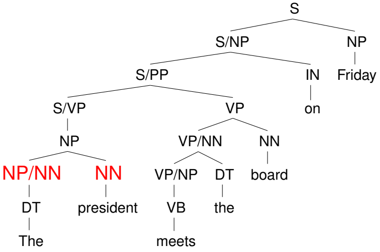


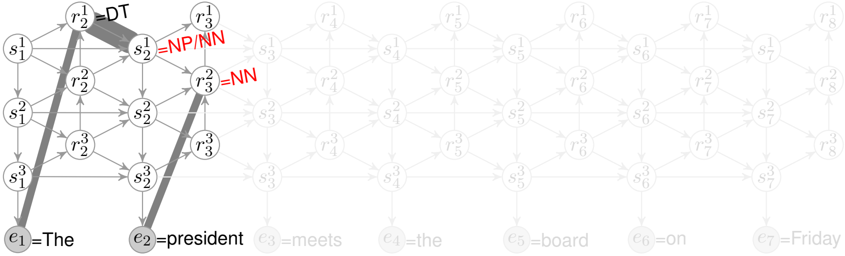
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to "Maximally Incremental" CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path



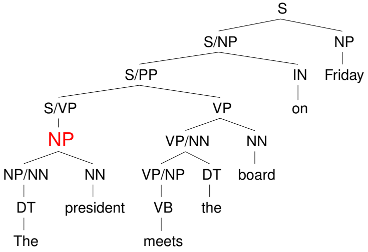


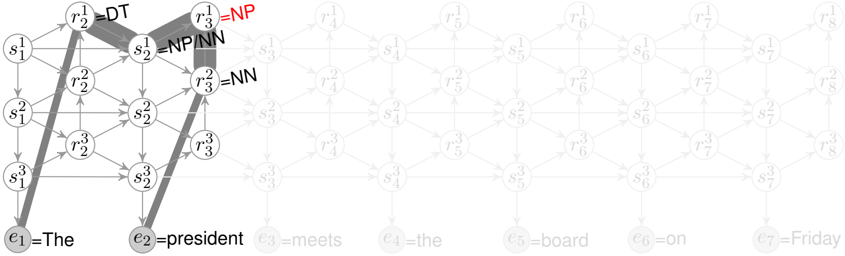
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to "Maximally Incremental" CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path



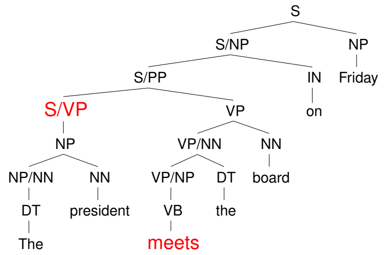


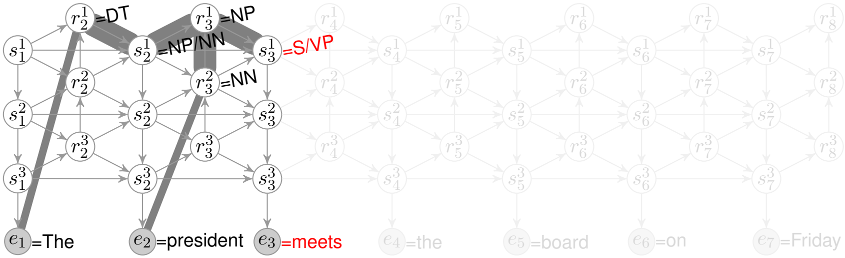
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree $\rightarrow$ Path



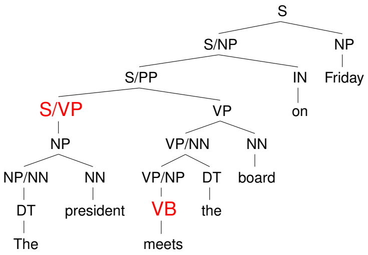


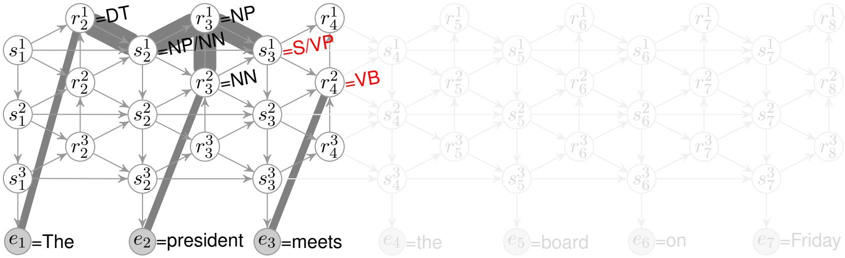
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree $\rightarrow$ Path



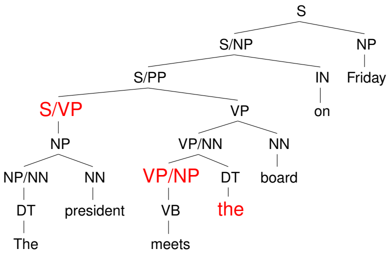


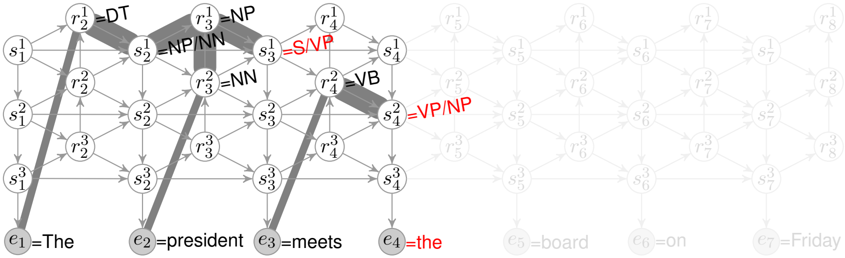
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path



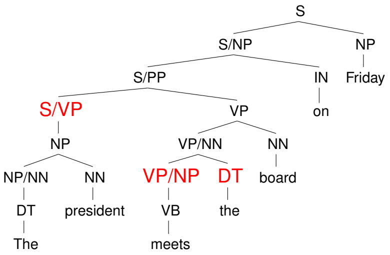


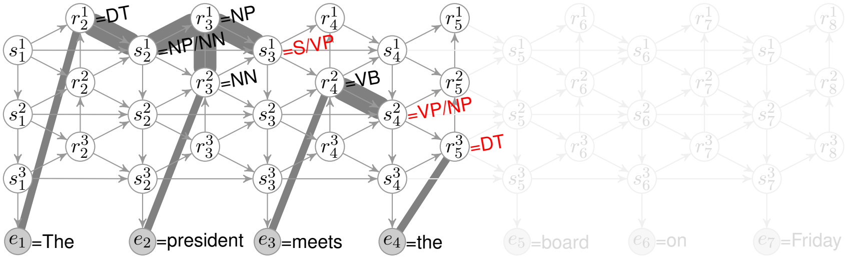
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree $\rightarrow$ Path



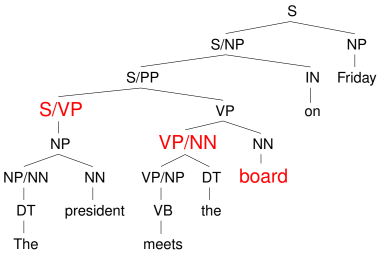


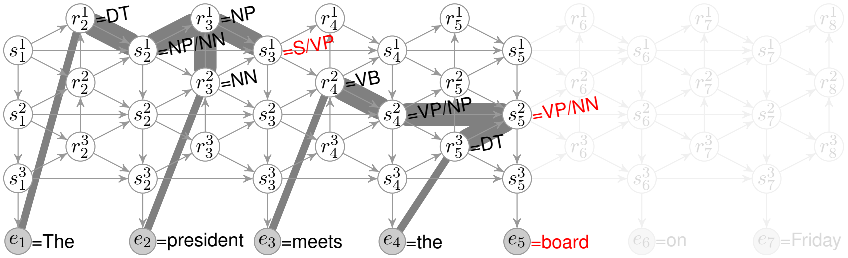
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to "Maximally Incremental" CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path



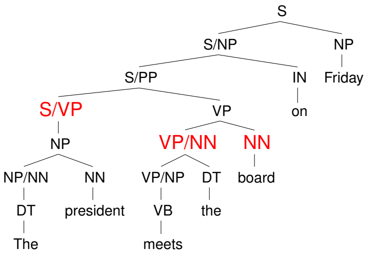


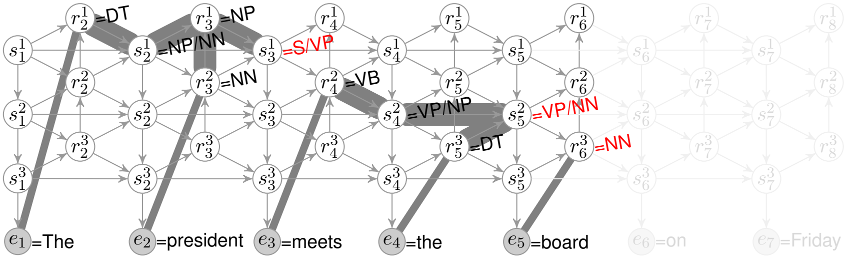
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to "Maximally Incremental" CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path



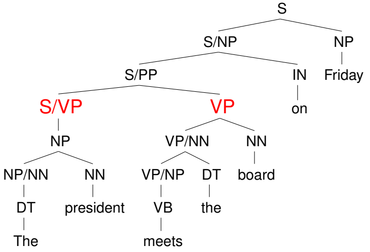


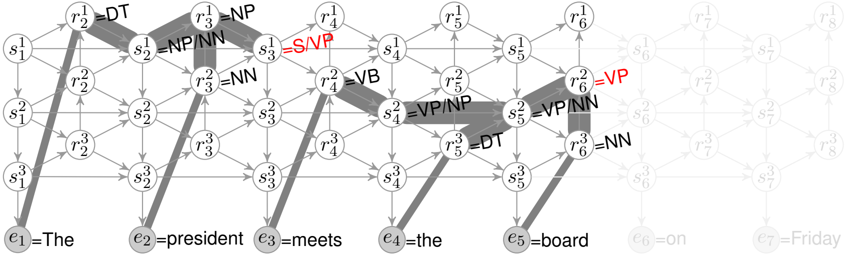
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to "Maximally Incremental" CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree $\rightarrow$ Path



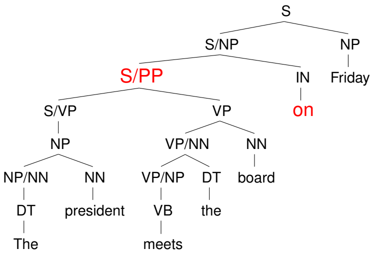


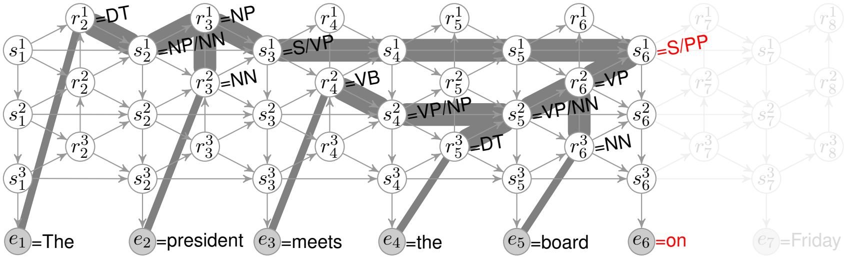
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to "Maximally Incremental" CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path



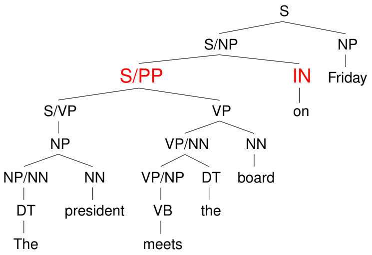


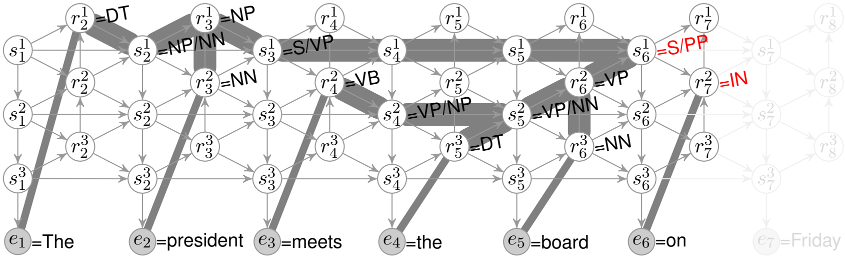
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path



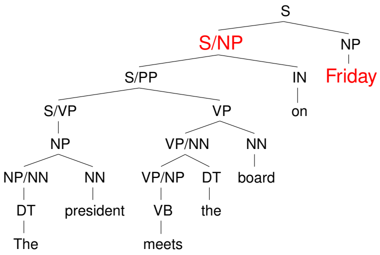


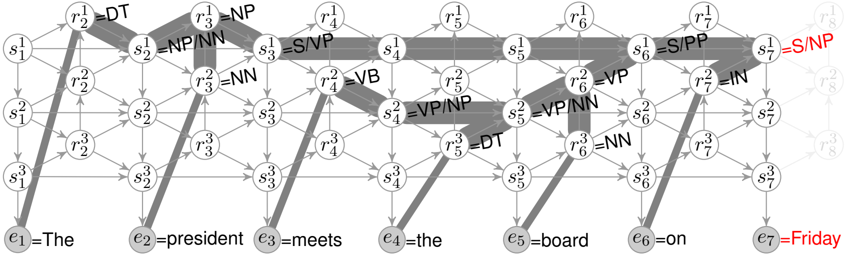
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to "Maximally Incremental" CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree $\rightarrow$ Path



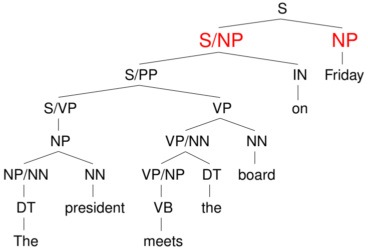


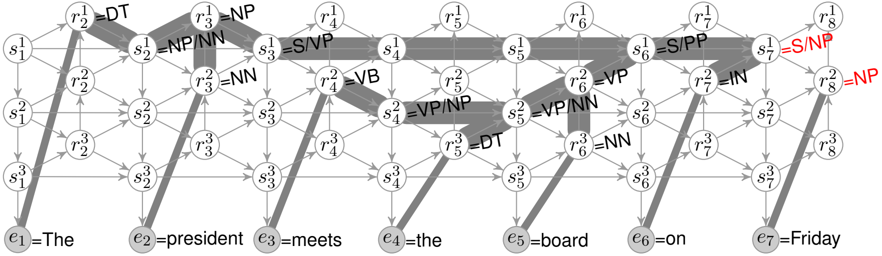
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path



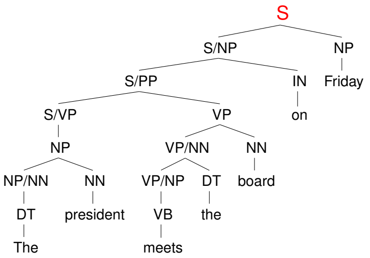


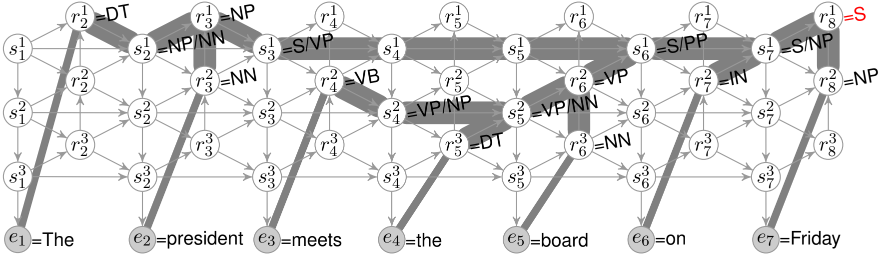
Incremental Parsing using HHMM (Schuler *et al.* 2010)

Analogous to “Maximally Incremental” CCG Parsing

Equivalent to Probabilistic Push-Down Automata

Isomorphic Tree → Path





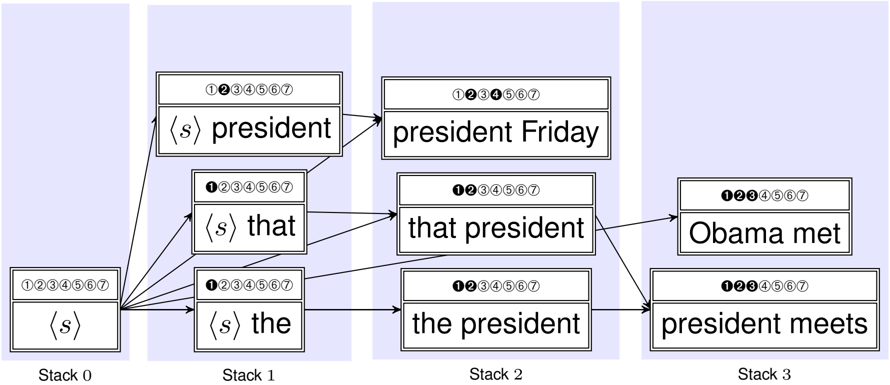
Phrase-Based Translation

Der Präsident trifft am Freitag den Vorstand

The president meets the board on Friday

Phrase-Based Translation

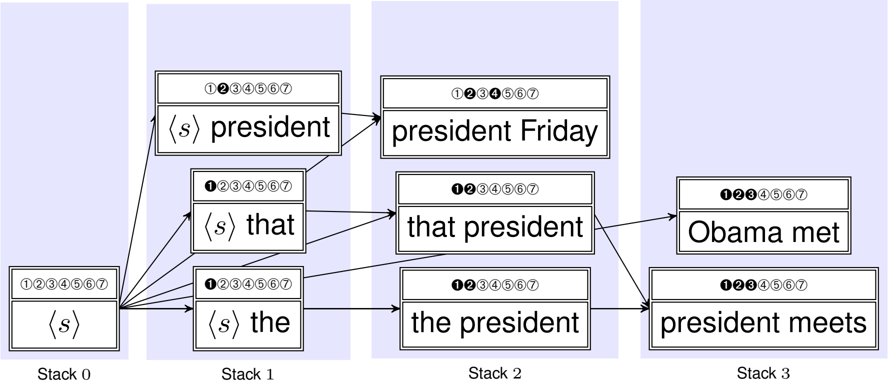
Der Präsident trifft am Freitag den Vorstand
The president meets the board on Friday



Phrase-Based Translation with Syntactic LM

Definition

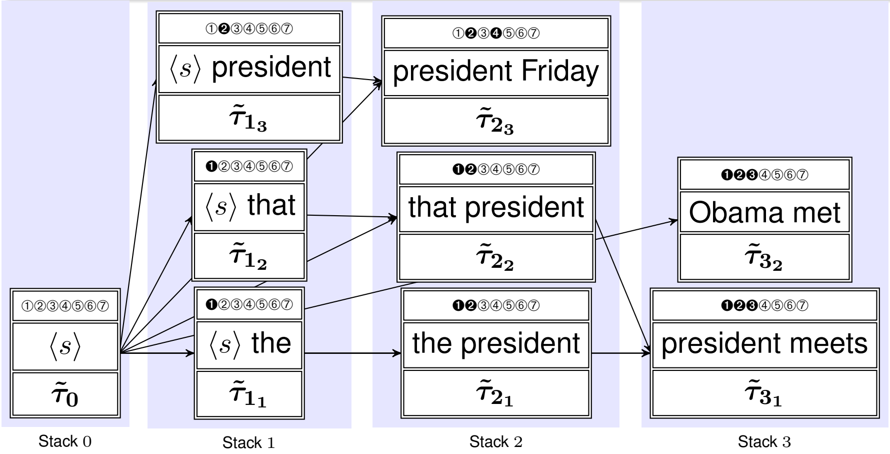
$\tilde{\boldsymbol{\tau}}_{t_h}$ represents parses of the partial translation at node h in stack t



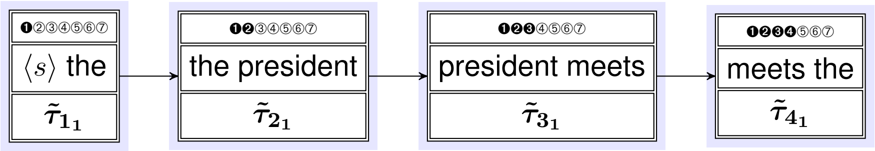
Phrase-Based Translation with Syntactic LM

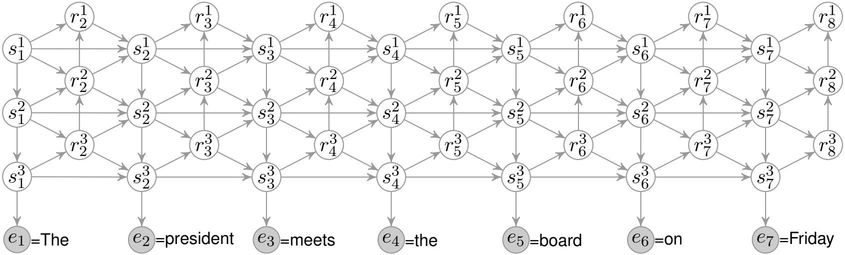
Definition

$\tilde{\tau}_{t_h}$ represents parses of the partial translation at node h in stack t

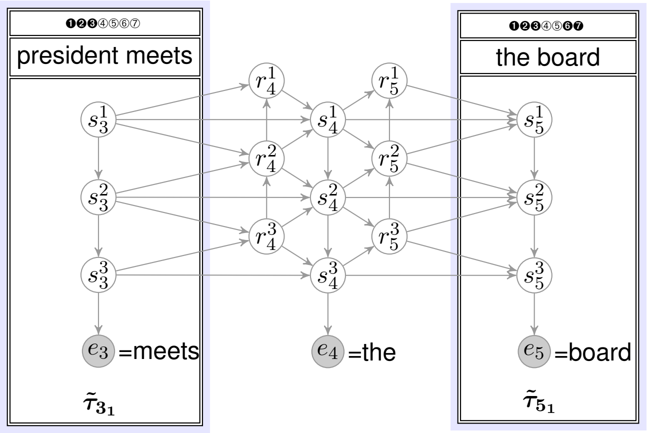


Integrate Parser into Phrase-based Decoder





Integrate Parser into Phrase-based Decoder

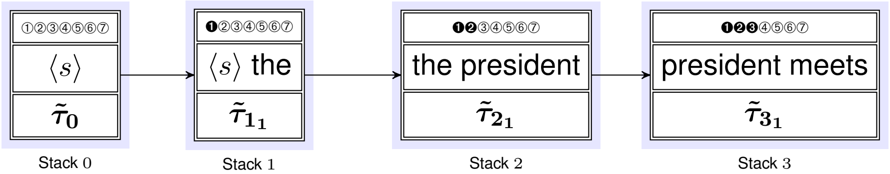


Direct Maximum Entropy Model of Translation

$$\hat{\boldsymbol{e}} = \underset{\boldsymbol{e}}{\operatorname{argmax}} \exp \sum_j \lambda_j h_j(\boldsymbol{e}, \boldsymbol{f})$$

$$\boldsymbol{\lambda} = \text{Set of } j \text{ feature weights}$$

$$\boldsymbol{h} = \begin{cases} \text{Phrase-based translation model} \\ n\text{-gram LM} \\ \text{Distortion model} \\ \vdots \\ \text{Syntactic LM } \mathsf{P}(\tilde{\boldsymbol{\tau}}_{t_h}) \end{cases}$$



Does an Incremental Syntactic LM Help Translation?

That's nice...

but will it make my BLEU score go up?

Language models trained on WSJ Treebank corpus

LM	In-domain Perplexity	Out-of-domain Perplexity
WSJ 5-gram LM	**232**	1262
WSJ Syntactic LM	385	**529**

Perplexity Results

Language models trained on WSJ Treebank corpus

LM	In-domain Perplexity	Out-of-domain Perplexity
WSJ 5-gram LM	**232**	1262
WSJ Syntactic LM	385	**529**
Interpolated WSJ 5-gram + WSJ SynLM	***209***	***225***

Perplexity Results

Language models trained on WSJ Treebank corpus
...and n-gram model for larger English Gigaword corpus.

LM	In-domain Perplexity	Out-of-domain Perplexity
WSJ 5-gram LM	**232**	1262
WSJ Syntactic LM	385	529
Interpolated WSJ 5-gram + WSJ SynLM	***209***	225

Gigaword 5-gram	258	**312**
Interpolated Gigaword 5-gram + WSJ SynLM	222	***123***

That's nice...

but will it make my BLEU score go up?

Does an Incremental Syntactic LM Help Translation?

Moses with LM(s)	BLEU
Using n-gram LM only	18.78
Using n-gram LM + Syntactic LM	**19.78**

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?

Moses with LM(s)	BLEU
Using n-gram LM only	18.78
Using n-gram LM + Syntactic LM	**19.78**

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`

This looks a lot like CCG

Our parser performs some CCG-style operations:

- Forward function application
  - NP/NN NN $\Rightarrow$ NP
- Type raising
  - NP $\Rightarrow$ S/VP
- Type raising in conjunction with forward function composition
  - DT $\Rightarrow$ NP/NN
  - VP/NP NP/NN $\Rightarrow$ VP/NN

Why not just use CCG?

- No probablistic 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

Sentence length	Moses	+SynLM beam=50	+SynLM beam=2000
10	0.2 sec	9 min	19 min
20	0.5 sec	20 min	43 min
30	0.9 sec	29 min	62 min
40	1.1 sec	35 min	76 min

- 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

Phrase-Based Translation with Syntactic LM

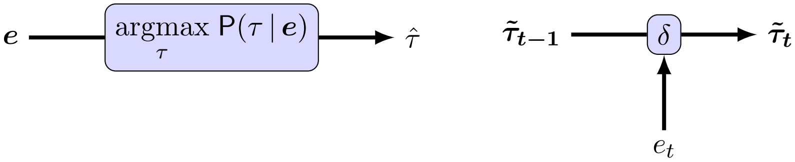
Definition

$\boldsymbol{e} \stackrel{\text{def}}{=}$ string of n target language words $e_1 \ldots e_n$

$\boldsymbol{e_t} \stackrel{\text{def}}{=}$ the first t words in $\boldsymbol{e}$, where $t \leq n$

$\boldsymbol{\tau_t} \stackrel{\text{def}}{=}$ set of all incremental parses of $\boldsymbol{e_t}$

$\boldsymbol{\tilde{\tau}_t} \stackrel{\text{def}}{=}$ subset of parses $\boldsymbol{\tau_t}$ that remain after parser pruning



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.