PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Neural Semantic Role Labeling: What works and what's next

or: What else can we do other than using 1000 LSTM layers :)

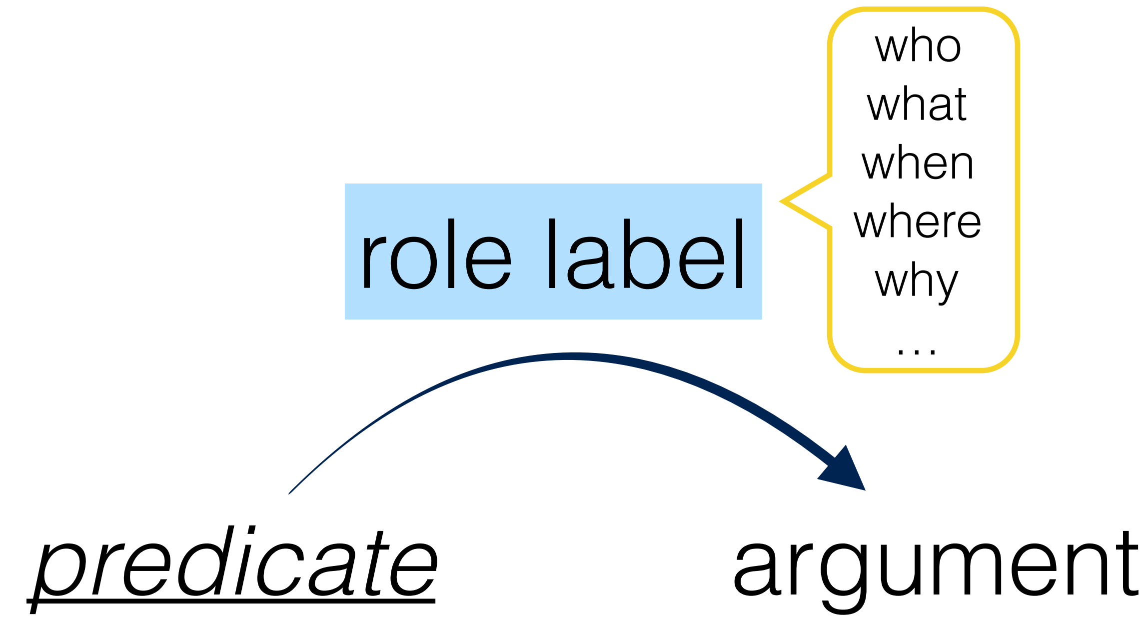
Luheng He[†], Kenton Lee[†], Mike Lewis [‡] and Luke Zettlemoyer[†*]

[†] Paul G. Allen School of Computer Science & Engineering, Univ. of Washington,
[‡] Facebook AI Research
[*] Allen Institute for Artificial Intelligence

Semantic Role Labeling (SRL)

Motivation

who did what to whom,
when and where



- Find out “who did what to whom” in text.
- Given predicate, identify arguments and label them.

Semantic Role Labeling (SRL)

Motivation

who did what to whom,
when and where

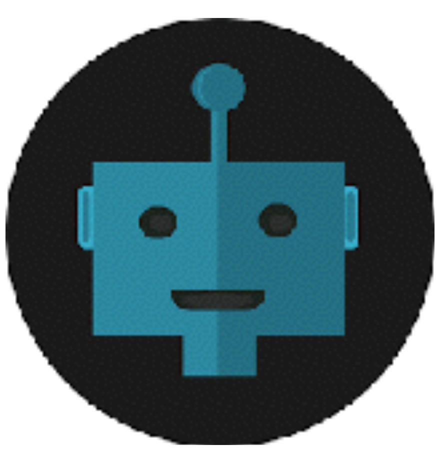
role label

who
what
when
where
why
…

predicate   argument

Applications

Question Answering



Information Extraction

Machine Translation

Semantic Role Labeling (SRL)

The robot *broke* my favorite mug with a wrench.

My mug *broke* into pieces immediately.

Semantic Role Labeling (SRL)

subj **v** **obj** **prep**

The robot *broke* my favorite mug with a wrench.

subj **v** **prep** **adv**

My mug *broke* into pieces immediately.

Semantic Role Labeling (SRL)

subj **v** **obj** **prep**

The robot *broke* my favorite mug with a wrench.

thing broken

subj **v** **prep** **adv**

My mug *broke* into pieces immediately.

thing broken

Semantic Role Labeling (SRL)

subj **v** **obj** **prep**

The robot *broke* my favorite mug with a wrench.

breaker | thing broken | instrument

subj **v** **prep** **adv**

My mug *broke* into pieces immediately.

thing broken | pieces (final state) | temporal

Semantic Role Labeling (SRL)

subj **v** **obj** **prep**

The robot *broke* my favorite mug with a wrench.

breaker thing broken instrument

subj **v** **prep** **adv**

My mug *broke* into pieces immediately.

thing broken pieces (final state) temporal

Frame: *break.01*

role	description
ARG0	breaker
ARG1	thing broken
ARG2	instrument
ARG3	pieces
ARG4	broken away from what?

Semantic Role Labeling (SRL)

subj **v** **obj** **prep**

The robot *broke* my favorite mug with a wrench.

breaker — ARG0; thing broken — ARG1; instrument — ARG2

subj **v** **prep** **adv**

My mug *broke* into pieces immediately.

thing broken — ARG1; pieces (final state) — ARG3; temporal — ARGM-TMP

Frame: *break.01*

role	description
ARG0	breaker
ARG1	thing broken
ARG2	instrument
ARG3	pieces
ARG4	broken away from what?

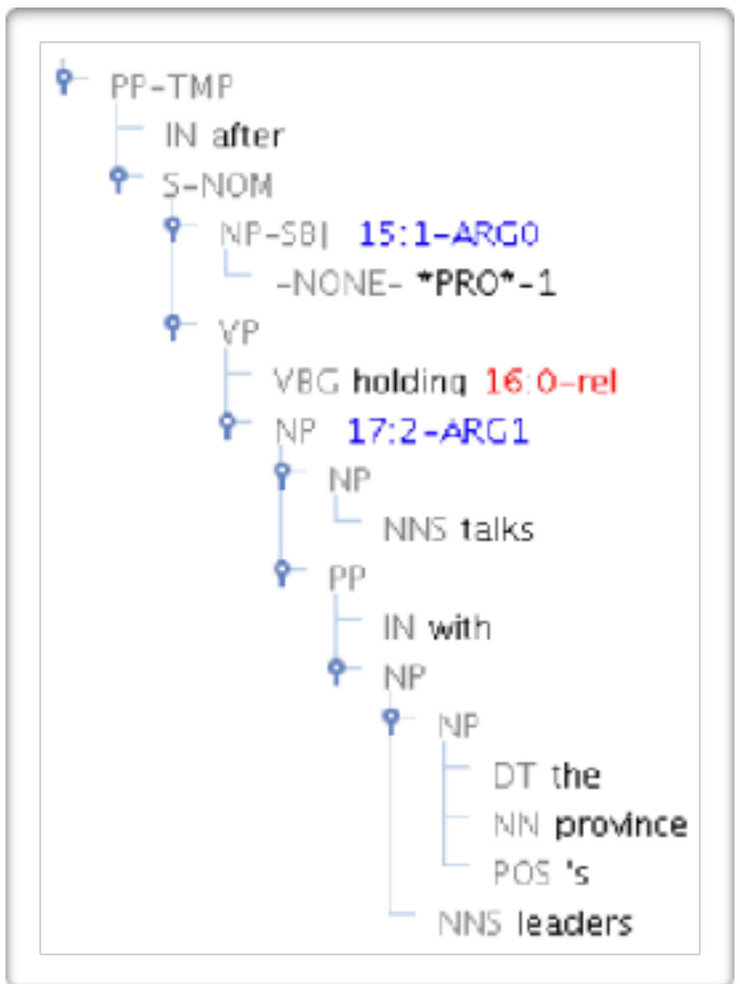
The Proposition Bank (PropBank)

Paul Kingsbury and Martha Palmer. From Treebank to PropBank. 2002

The Proposition Bank (PropBank)

Paul Kingsbury and Martha Palmer. From Treebank to PropBank. 2002

Annotated on top of the Penn Treebank Syntax

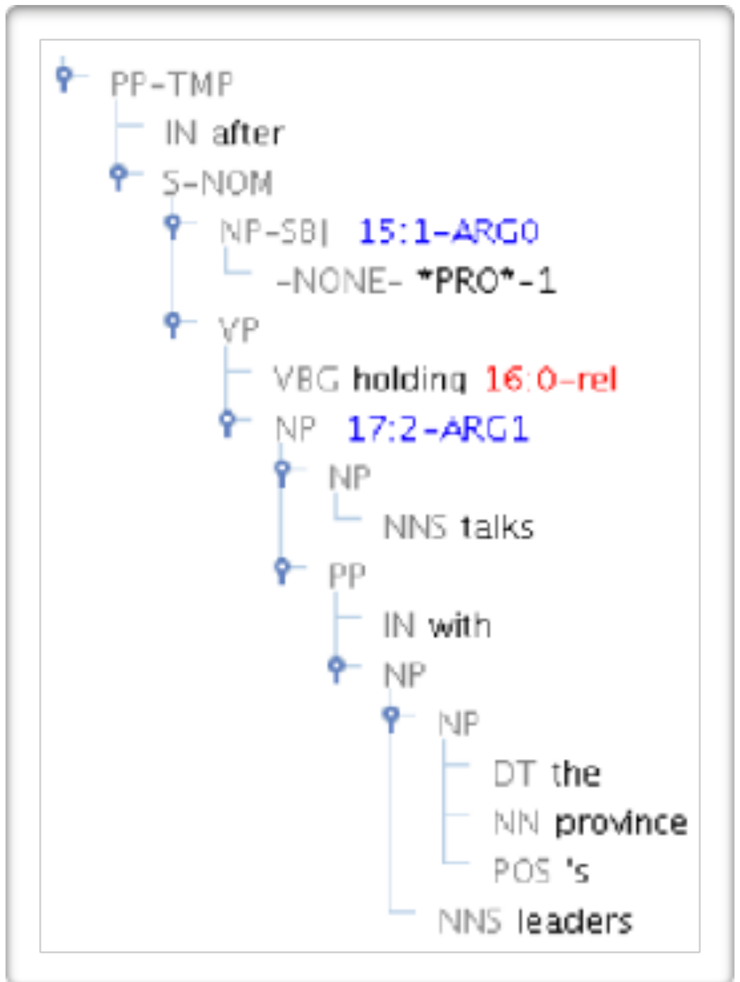


PropBank Annotation Guidelines, Bonial et al., 2010

The Proposition Bank (PropBank)

Paul Kingsbury and Martha Palmer. From Treebank to PropBank. 2002

Annotated on top of the Penn Treebank Syntax



PropBank Annotation Guidelines, Bonial et al., 2010

Core roles:
Verb-specific roles (ARG0-ARG5) defined in frame files

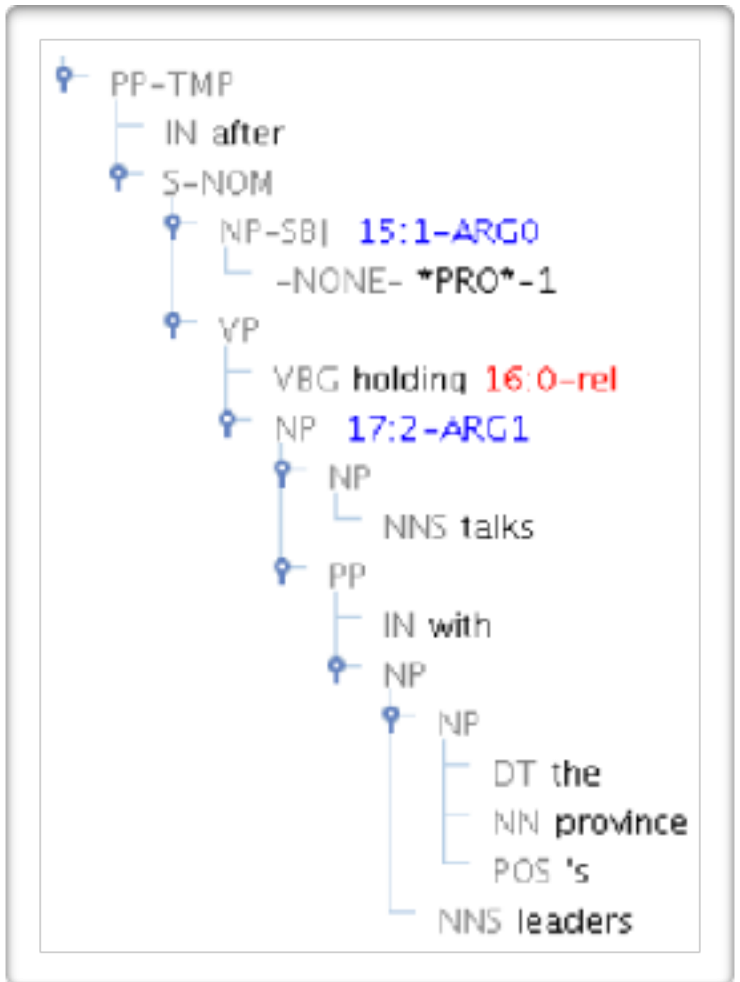
Frame: *break.01*

role	description
ARG0	breaker
ARG1	thing broken
ARG2	instrument
ARG3	pieces
ARG4	broken away from what?

The Proposition Bank (PropBank)

Paul Kingsbury and Martha Palmer. From Treebank to PropBank. 2002

Annotated on top of the Penn Treebank Syntax



PropBank Annotation Guidelines, Bonial et al., 2010

Core roles:
Verb-specific roles (ARG0-ARG5) defined in frame files

Frame: break.01

role	description
ARG0	breaker
ARG1	thing broken
ARG2	instrument

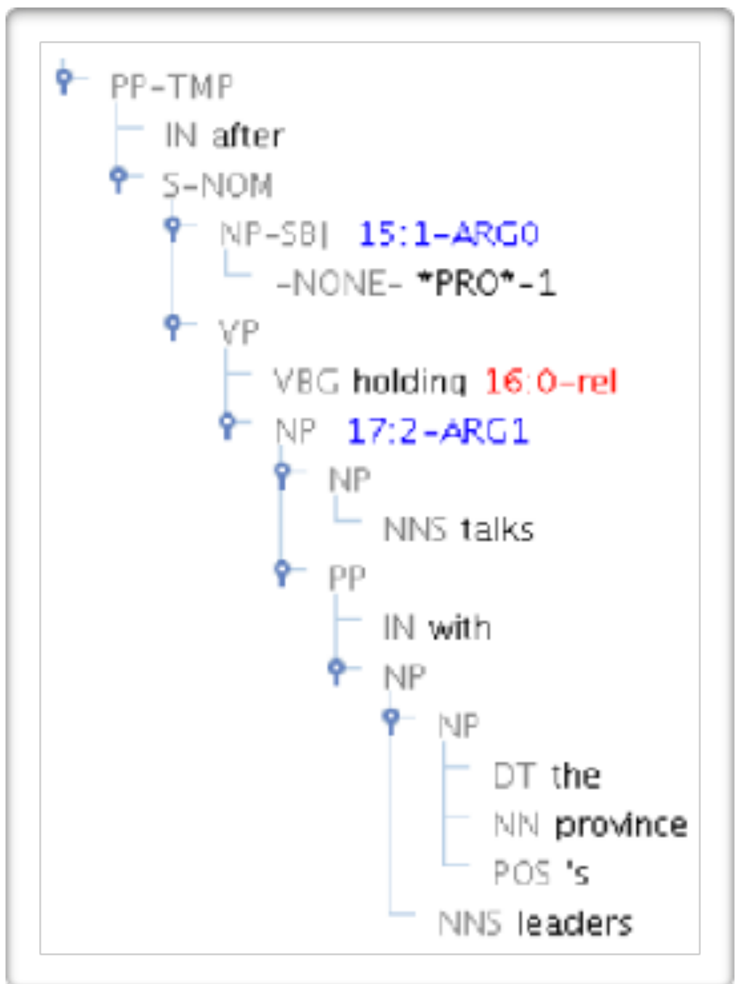
Frame: buy.01

role	description
ARG0	buyer
ARG1	thing bough
ARG2	seller
ARG3	price paid
ARG4	benefactive

The Proposition Bank (PropBank)

Paul Kingsbury and Martha Palmer. From Treebank to PropBank. 2002

Annotated on top of the Penn Treebank Syntax



PropBank Annotation Guidelines, Bonial et al., 2010

Core roles:
Verb-specific roles (ARG0-ARG5) defined in frame files

Frame: *break.01*

role	description
ARG0	breaker
ARG1	thing broken
ARG2	instrument

Frame: *buy.01*

role	description
ARG0	buyer
ARG1	thing bough
ARG2	seller
ARG3	price paid
ARG4	benefactive

Adjunct roles:
(ARGM-) shared across verbs

role	description
TMP	temporal
LOC	location
MNR	manner
DIR	direction
CAU	cause
PRP	purpose
…	

SRL is a hard problem …

SRL is a hard problem …

- Over 10 years, F1 on the PropBank test set:
  79.4 (Punyakanok 2005) — **80.3** (FitzGerald 2015)

SRL is a hard problem …

- Over 10 years, F1 on the PropBank test set:
  79.4 (Punyakanok 2005) — **80.3** (FitzGerald 2015)
- Many interesting challenges:
  Syntactic alternation
  Prepositional phrase attachment
  Long-range dependencies and common sense

Syntactic Alternation

The robot *plays* piano.

ARG0 — player

ARG2 — instrument

The cafe is *playing* my favorite song.

ARG0 — player

ARG1 — thing performed

The music *plays* softly.

ARG1 — thing performed

ARGM-MNR

Syntactic Alternation

subjects

The robot *plays* piano.

ARG0 — player

ARG2 — instrument

The cafe is *playing* my favorite song.

ARG0 — player

ARG1 — thing performed

The music *plays* softly.

ARG1 — thing performed

ARGM-MNR

Syntactic Alternation

The robot *plays* piano.

ARG0 — player

ARG2 — instrument

objects

subjects

The cafe is *playing* my favorite song.

ARG0 — player

ARG1 — thing performed

The music *plays* softly.

ARG1 — thing performed

ARGM-MNR

Prepositional Phrase (PP) Attachment

I *eat* [pasta] [with delight].

ARG0 eater | ARG1 meal | ARGM-MNR manner

Prepositional Phrase (PP) Attachment

I *eat* [pasta] [with delight].

ARG0 ARG1 ARGM-MNR

eater meal manner

I *eat* [pasta with broccoli].

ARG0 ARG1

eater meal

Long-range Dependencies

We *flew* to Chicago.

ARG1 ARGM-GOL

passenger goal

Long-range Dependencies

We *flew* to Chicago.

ARG1 — passenger

ARGM-GOL — goal

We remember the nice view *flying* to Chicago.

ARG1 — passenger

ARGM-GOL — goal

Long-range Dependencies

We *flew* to Chicago.

ARG1 — passenger

ARGM-GOL — goal

We remember the nice view *flying* to Chicago.

ARG1 — passenger

ARGM-GOL — goal

We remember John and Mary *flying* to Chicago.

ARG1 — passenger

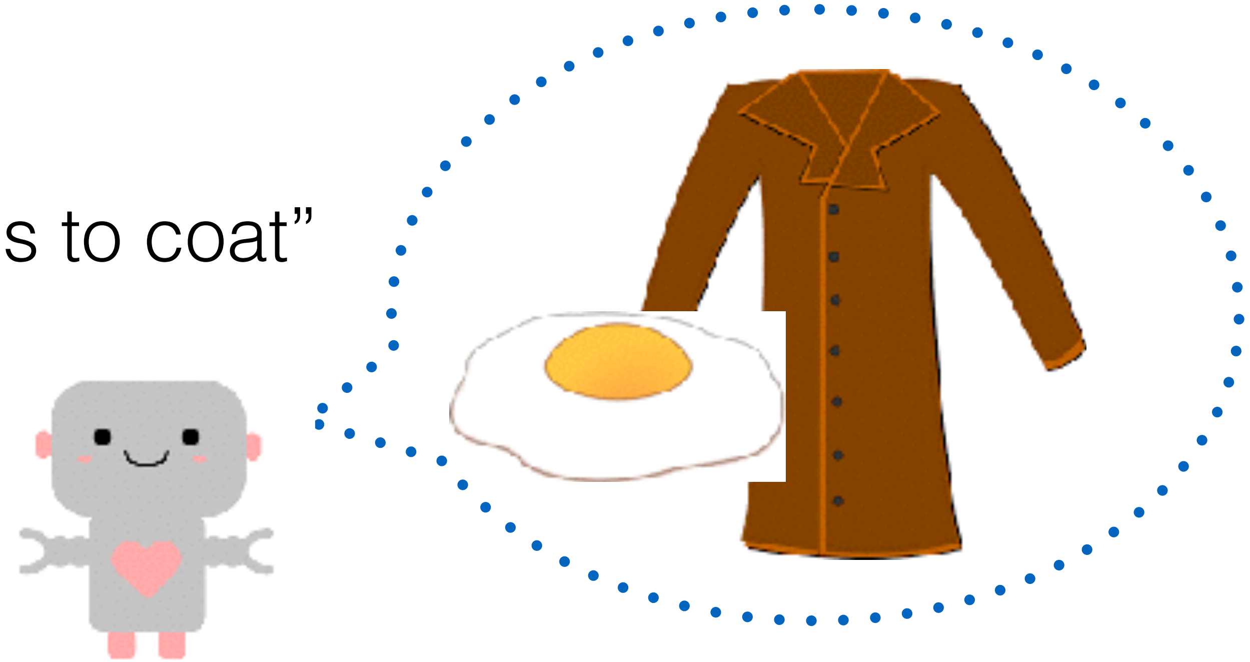
ARGM-GOL — goal

SRL is even harder for out-domain data …

"Dip chicken breasts into eggs to coat"

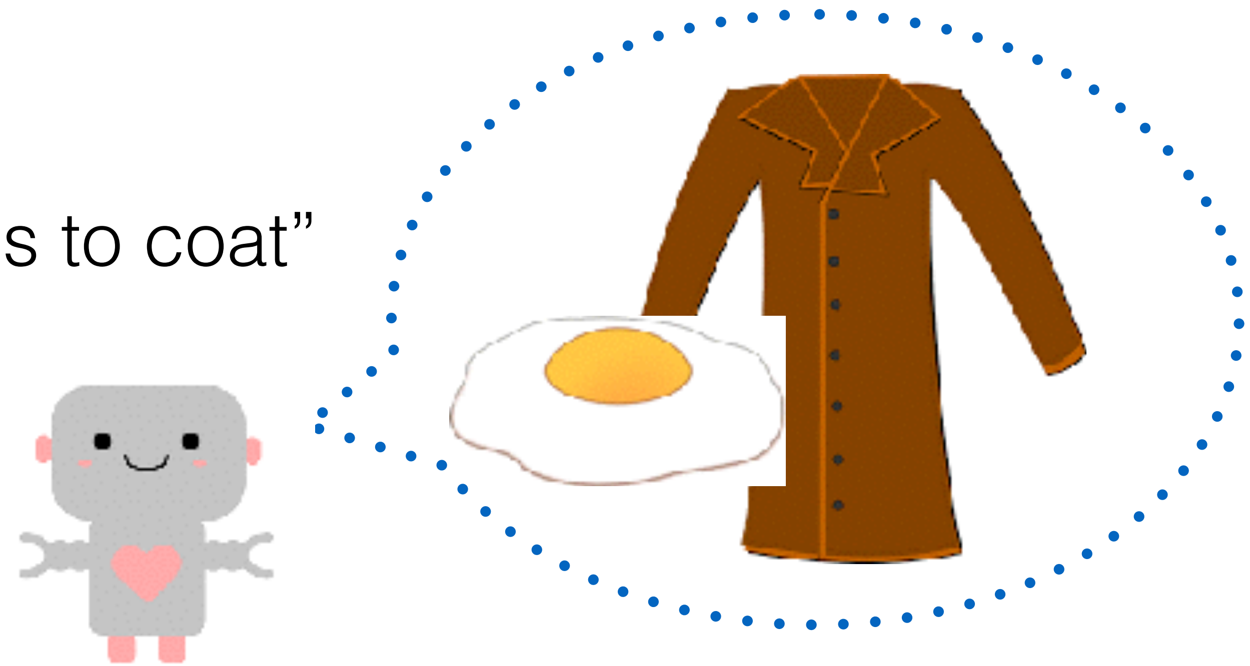
SRL is even harder for out-domain data …

"Dip chicken breasts into eggs to coat"



SRL is even harder for out-domain data …

"Dip chicken breasts into eggs to coat"



Active, Ser133-phosphorylated CREB effects transcription of CRE-dependent genes via interaction with the 265-kDa …

Long-term Plan for Improving SRL

Long-term Plan for Improving SRL

Step 1: Collect more data for SRL

— Question-Answer Driven Semantic Role Labeling (QA-SRL)

— Human-in-the-Loop Parsing

Long-term Plan for Improving SRL

Step 1: Collect more data for SRL

— Question-Answer Driven Semantic Role Labeling (QA-SRL)

— Human-in-the-Loop Parsing

Step 2: Build accurate SRL model

— Neural Semantic Role Labeling (for PropBank SRL)

Long-term Plan for Improving SRL

Step 1: Collect more data for SRL

— Question-Answer Driven Semantic Role Labeling (QA-SRL)

— Human-in-the-Loop Parsing

Step 2: Build accurate SRL model

— Neural Semantic Role Labeling (for PropBank SRL)

Step 3: SRL system for many domains

— *Future work …*

Long-term Plan for Improving SRL

Step 1: Collect more data for SRL

— Question-Answer Driven Semantic Role Labeling (QA-SRL)

— Human-in-the-Loop Parsing

Step 2: Build accurate SRL model

— Neural Semantic Role Labeling (for PropBank SRL)

Step 3: SRL system for many domains

— *Future work …*

First Step: Collect more (cheaper) SRL Data

First Step: Collect more (cheaper) SRL Data

Intuition: *Anyone who understands the meaning of a sentence should be able provide annotation for SRL*.

First Step: Collect more (cheaper) SRL Data

Intuition: *Anyone who understands the meaning of a sentence should be able provide annotation for SRL.*

Challenge: Complicated annotation process of traditional SRL.

First Step: Collect more (cheaper) SRL Data

Intuition: *Anyone who understands the meaning of a sentence should be able provide annotation for SRL*.

Challenge: Complicated annotation process of traditional SRL.

Solution: Design a simpler annotation scheme!

Question-Answer Driven SRL (QA-SRL)

Given sentence and a verb:

Last month, we saw the Grand Canyon *flying* to Chicago.

Question-Answer Driven SRL (QA-SRL)

Given sentence and a verb:

Last month, we saw the Grand Canyon *flying* to Chicago.

Step 1: Ask a question about the verb:

Who was flying?

Question-Answer Driven SRL (QA-SRL)

Given sentence and a verb:

Last month, we saw the Grand Canyon *flying* to Chicago.

Step 1: Ask a question about the verb:

Who was flying?

Step 2: Answer with words in the sentence:

we

Question-Answer Driven SRL (QA-SRL)

Given sentence and a verb:

Last month, we saw the Grand Canyon *flying* to Chicago.

Step 1: Ask a question about the verb:

Who was flying?

Step 2: Answer with words in the sentence:

we

Step 3: Repeat, write as many Q/A pairs as possible …

Question-Answer Driven SRL (QA-SRL)

Given sentence and a verb:

Last month, we saw the Grand Canyon *flying* to Chicago.

Step 1: Ask a question about the verb:

Step 2: Answer with words in the sentence:

Who was flying?	we

Step 3: Repeat, write as many Q/A pairs as possible …

Where did someone fly to?	Chicago
When did someone fly?	Last month

Question-Answer Driven SRL (QA-SRL)

Given sentence and a verb:

Last month, we saw the Grand Canyon *flying* to Chicago.

Step 1: Ask a question about the verb:

Step 2: Answer with words in the sentence:

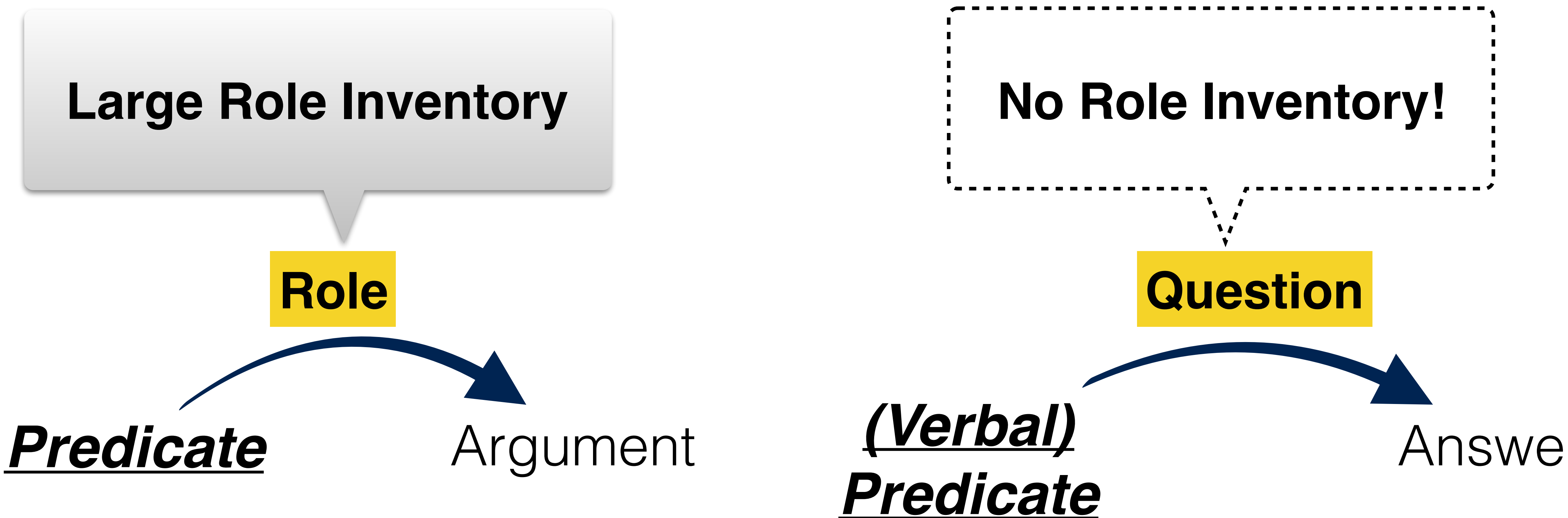
Question	Answer
Who was flying?	we

Step 3: Repeat, write as many Q/A pairs as possible …

Question	Answer
Where did someone fly to?	Chicago
When did someone fly?	Last month

Stop until all Q/A pairs are exhausted.

Comparing QA-SRL to PropBank



Traditional SRL
(PropBank)

Question-Answer Driven SRL
(QA-SRL)

Question-Answer Driven SRL (QA-SRL)

QA-SRL

Who was flying?

Where did someone fly to?

Last month, we saw the Grand Canyon *<u>flying</u>* to Chicago.

When did someone fly?

Question-Answer Driven SRL (QA-SRL)

QA-SRL **PropBank SRL**

ARG1

Who was flying?

ARGM-GOL

Where did someone fly to?

Last month, we saw the Grand Canyon *flying* to Chicago.

When did someone fly?

ARGM-TMP

Question-Answer Driven SRL (QA-SRL)

QA-SRL PropBank SRL

ARG1

Who was flying?

ARGM-GOL

Where did someone fly to?

Last month, we saw the Grand Canyon *flying* to Chicago.

When did someone fly?

ARGM-TMP

Non-expert annotated QA-SRL has about **80% agreement** with PropBank.

Long-term Plan for Improving SRL

Step 1: Collect more data for SRL

— Question-Answer Driven Semantic Role Labeling (QA-SRL)

— Human-in-the-Loop Parsing

Step 2: Build accurate SRL model

— Neural Semantic Role Labeling (for PropBank SRL)

Step 3: SRL system for many domains

— *Future work …*

Long-term Plan for Improving SRL

Step 1: Collect more data for SRL

— Question-Answer Driven Semantic Role Labeling (QA-SRL)

— Human-in-the-Loop Parsing

✓

Step 2: Build accurate SRL model

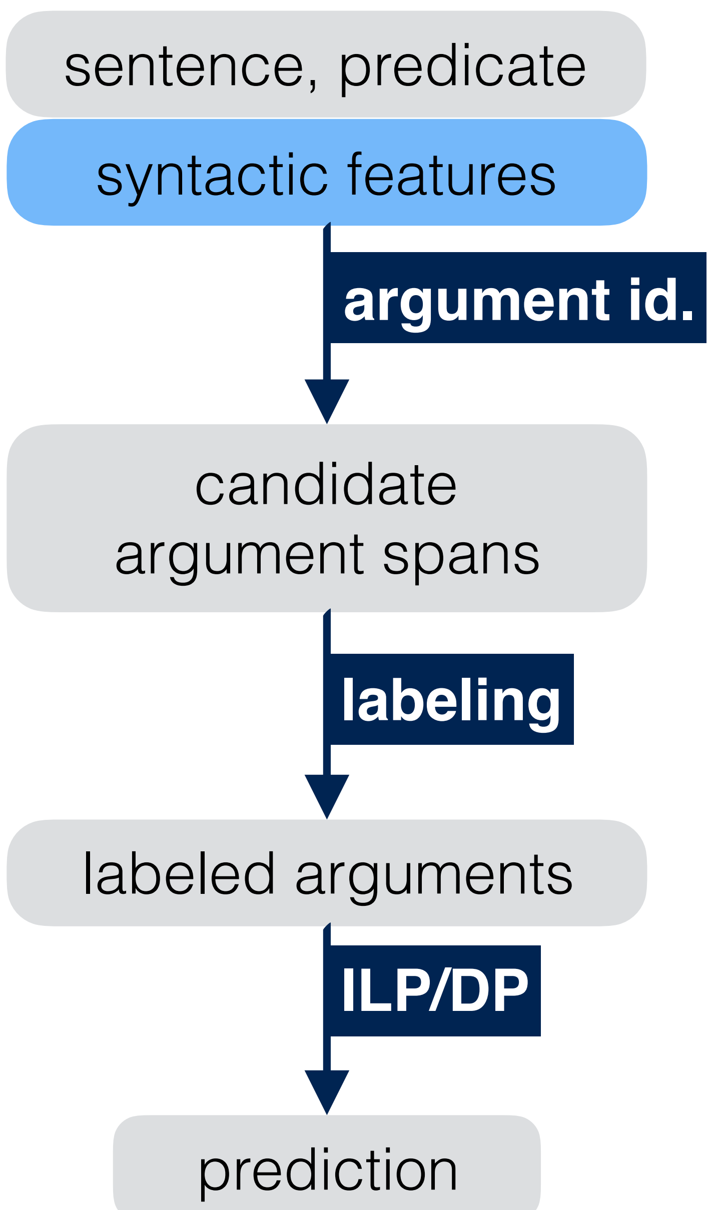
— Neural Semantic Role Labeling (for PropBank SRL)

Step 3: SRL system for many domains

— *Future work …*

SRL Systems

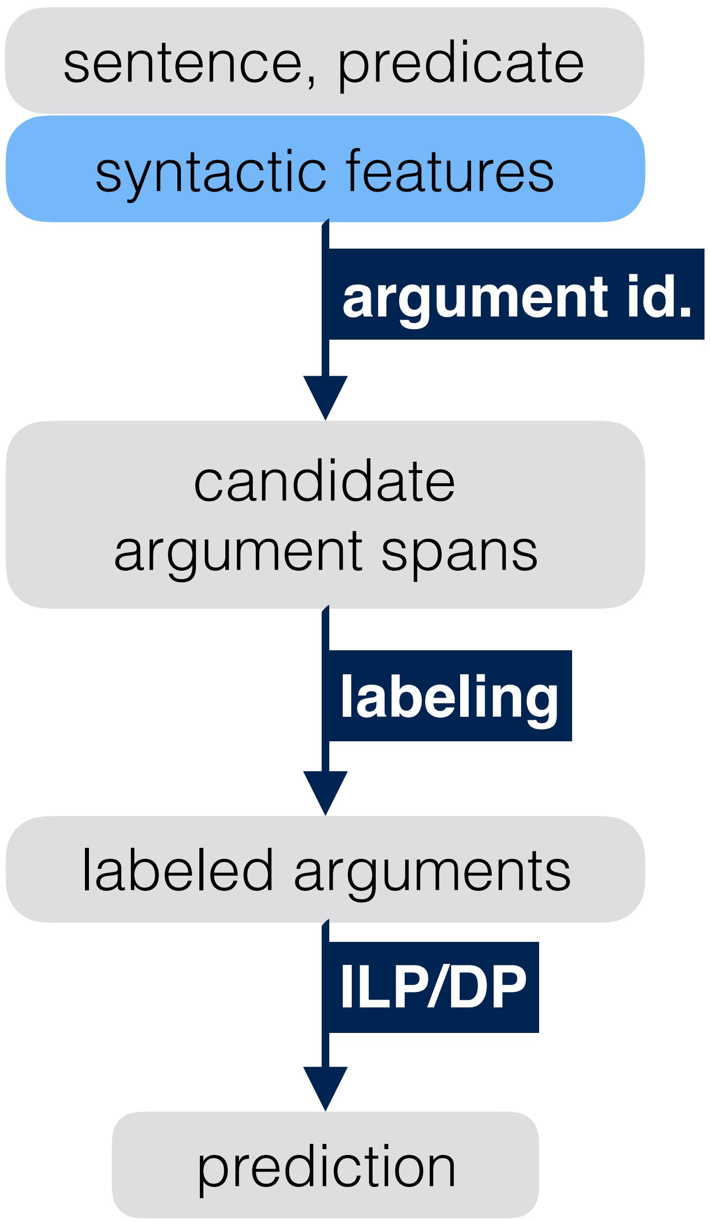
Pipeline Systems



Punyakanok et al., 2008
Täckström et al., 2015
FitzGerald et al., 2015

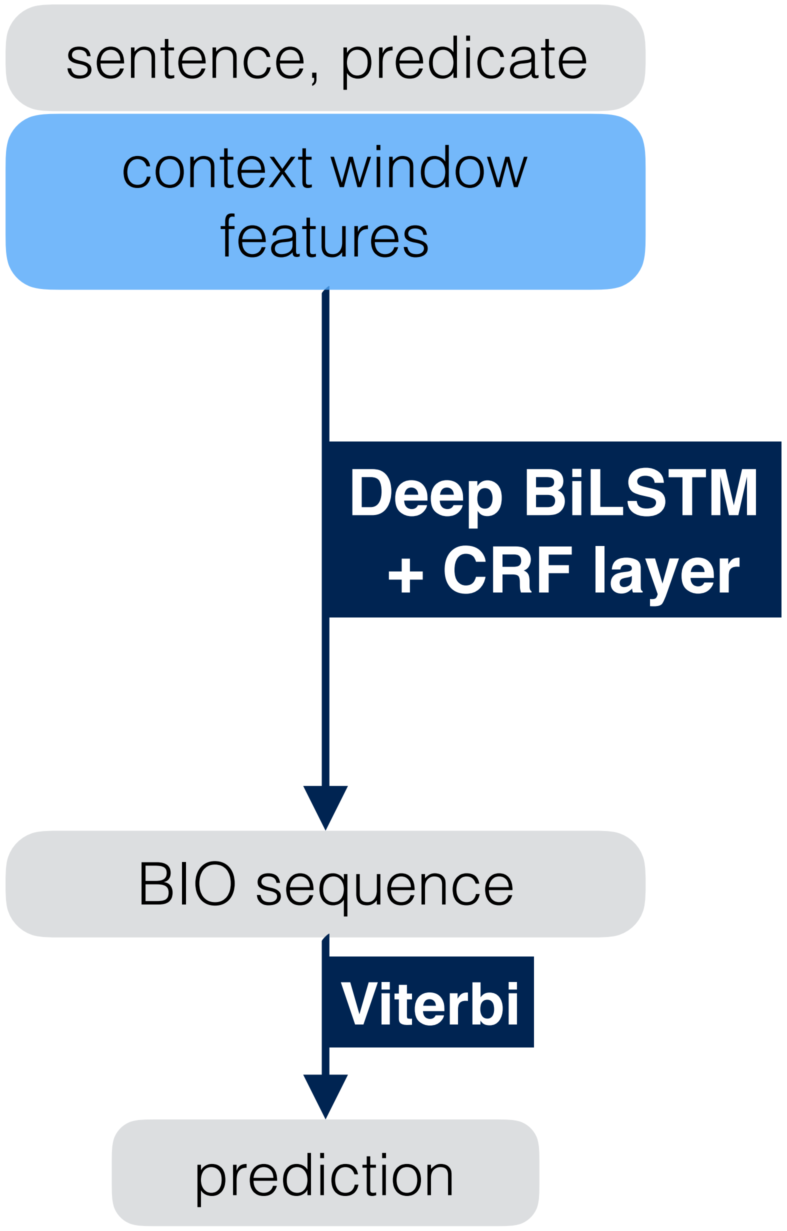
SRL Systems

Pipeline Systems



Punyakanok et al., 2008
Täckström et al., 2015
FitzGerald et al., 2015

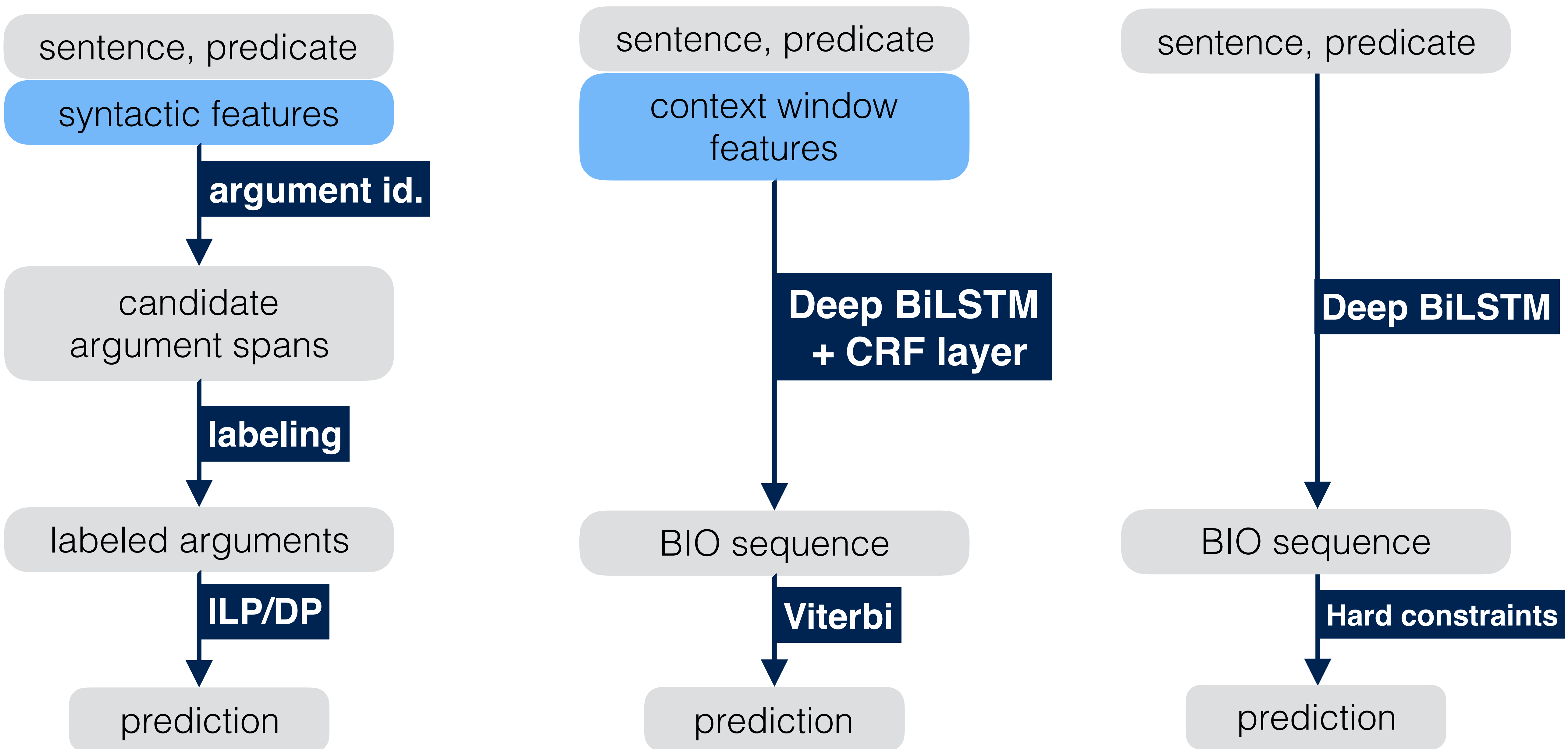
End-to-end Systems



Collobert et al., 2011
Zhou and Xu, 2015
Wang et. al, 2015

SRL Systems

Pipeline Systems | End-to-end Systems | ***This work**



Punyakanok et al., 2008
Täckström et al., 2015
FitzGerald et al., 2015

Collobert et al., 2011
Zhou and Xu, 2015
Wang et. al, 2015

SRL as BIO Tagging Problem

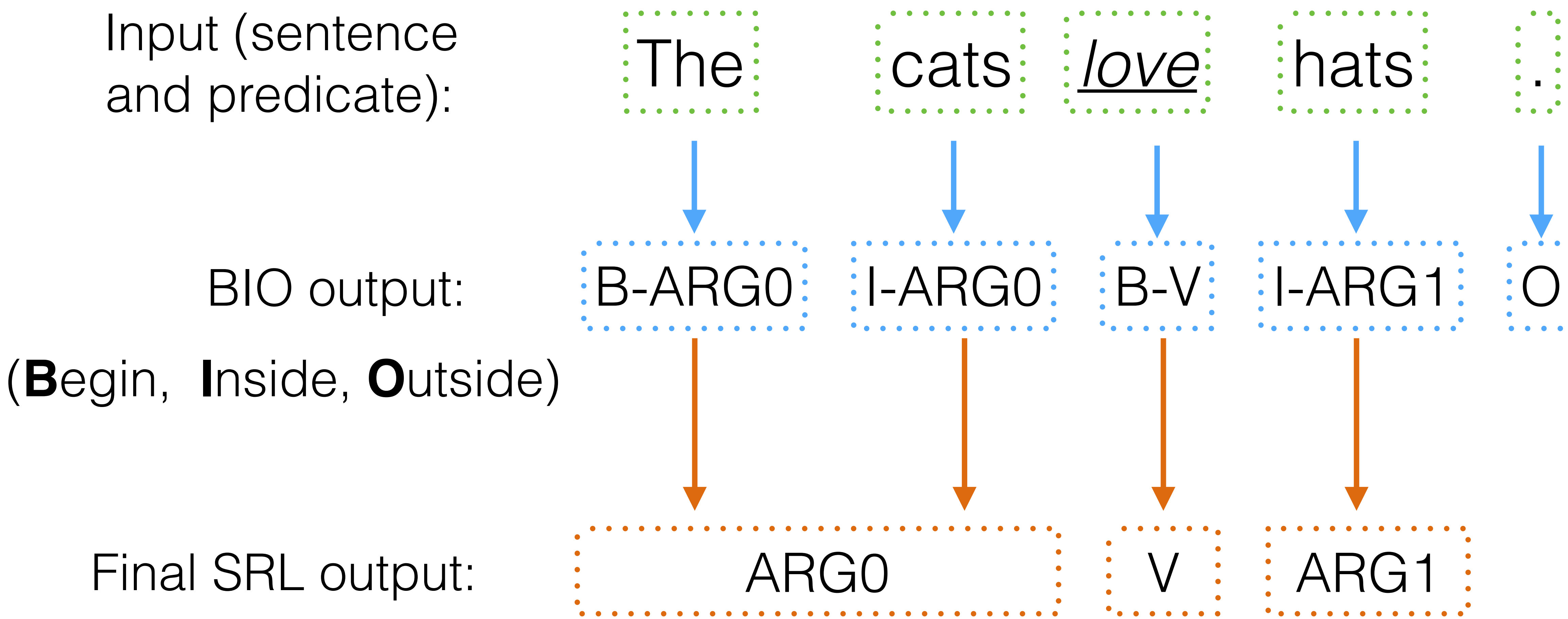
Input (sentence and predicate):	The	cats	*love*	hats	.

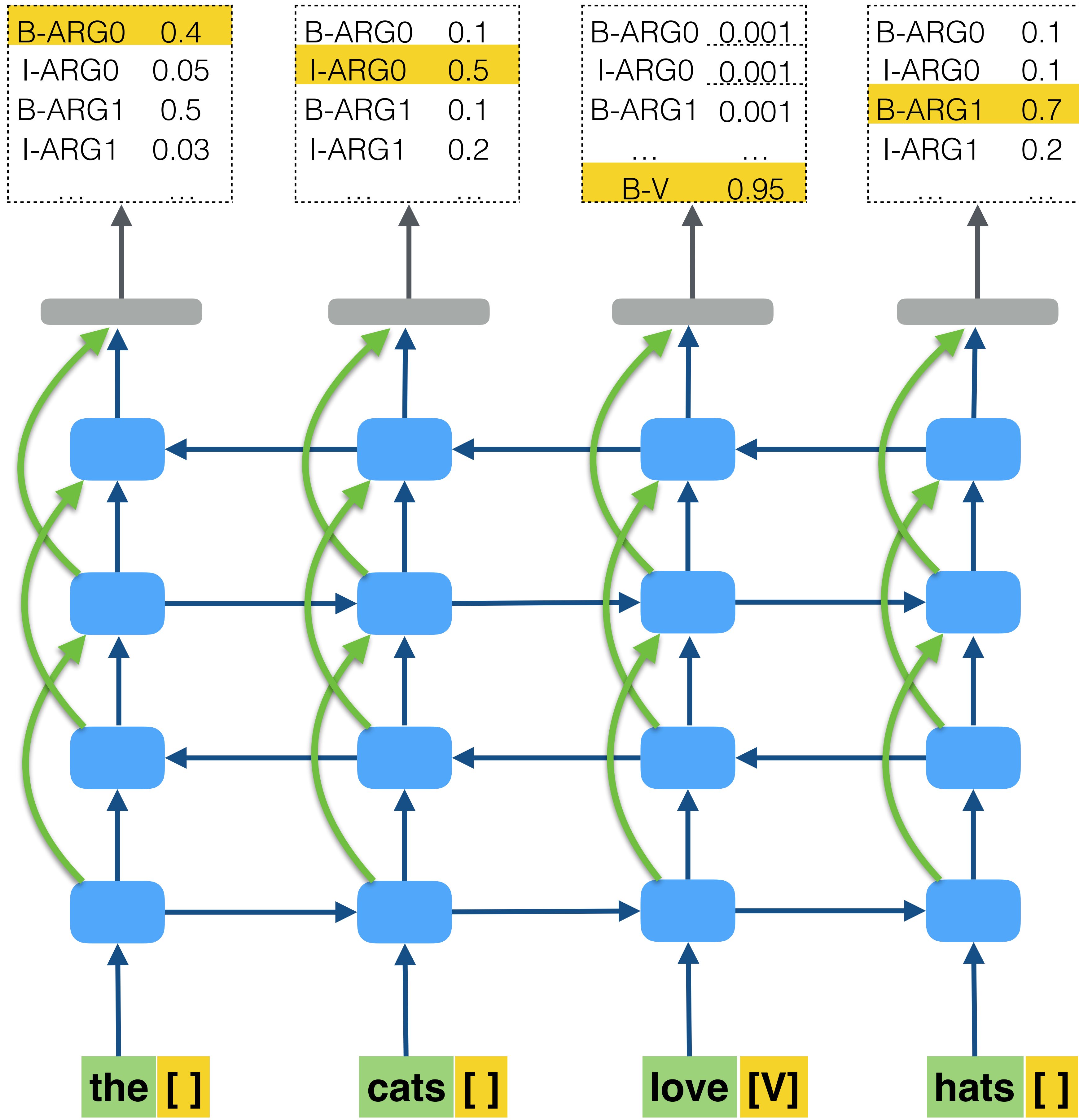
SRL as BIO Tagging Problem

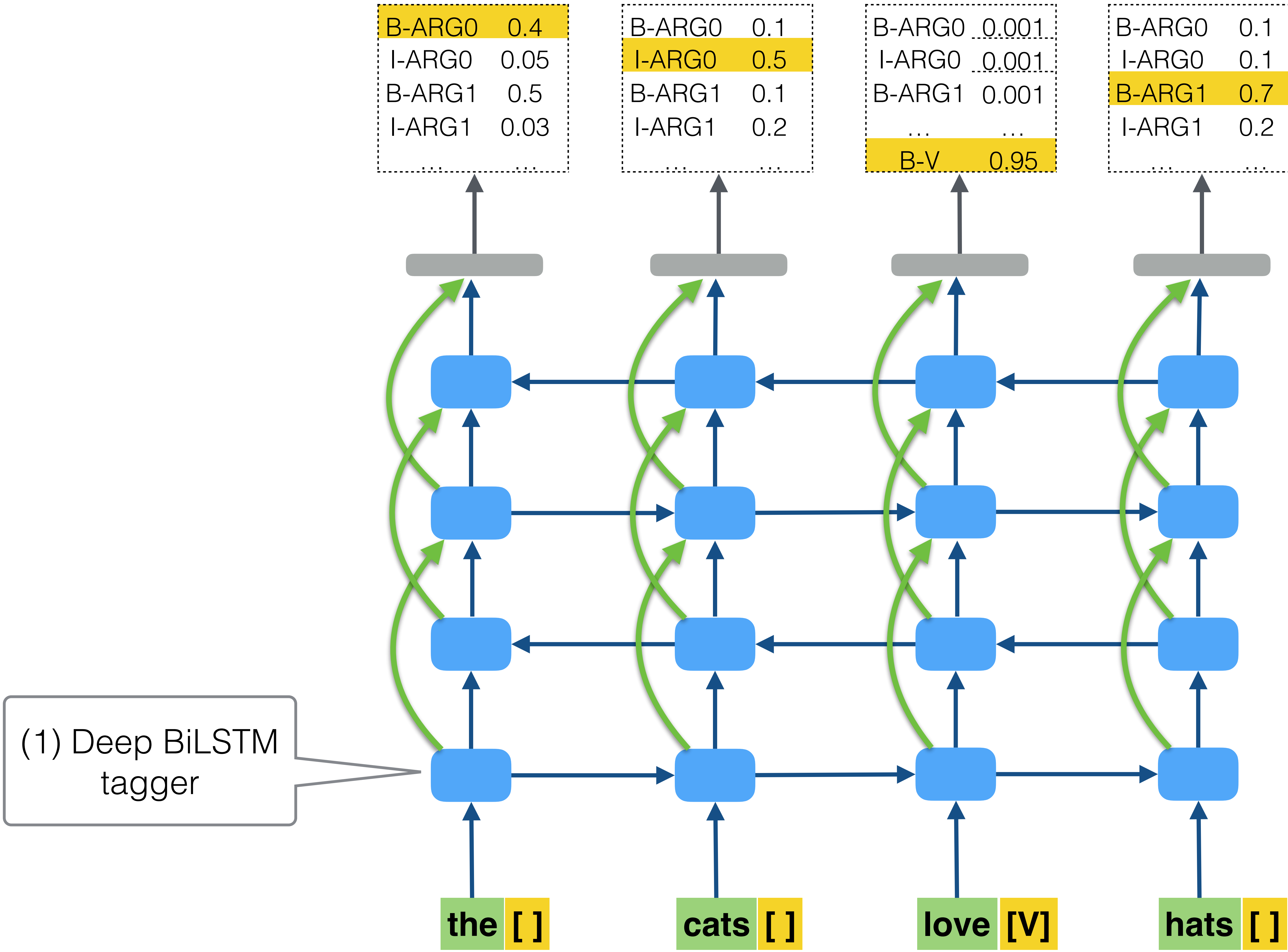
Input (sentence and predicate):	The	cats	*<u>love</u>*	hats	.
BIO output:	B-ARG0	I-ARG0	B-V	I-ARG1	O

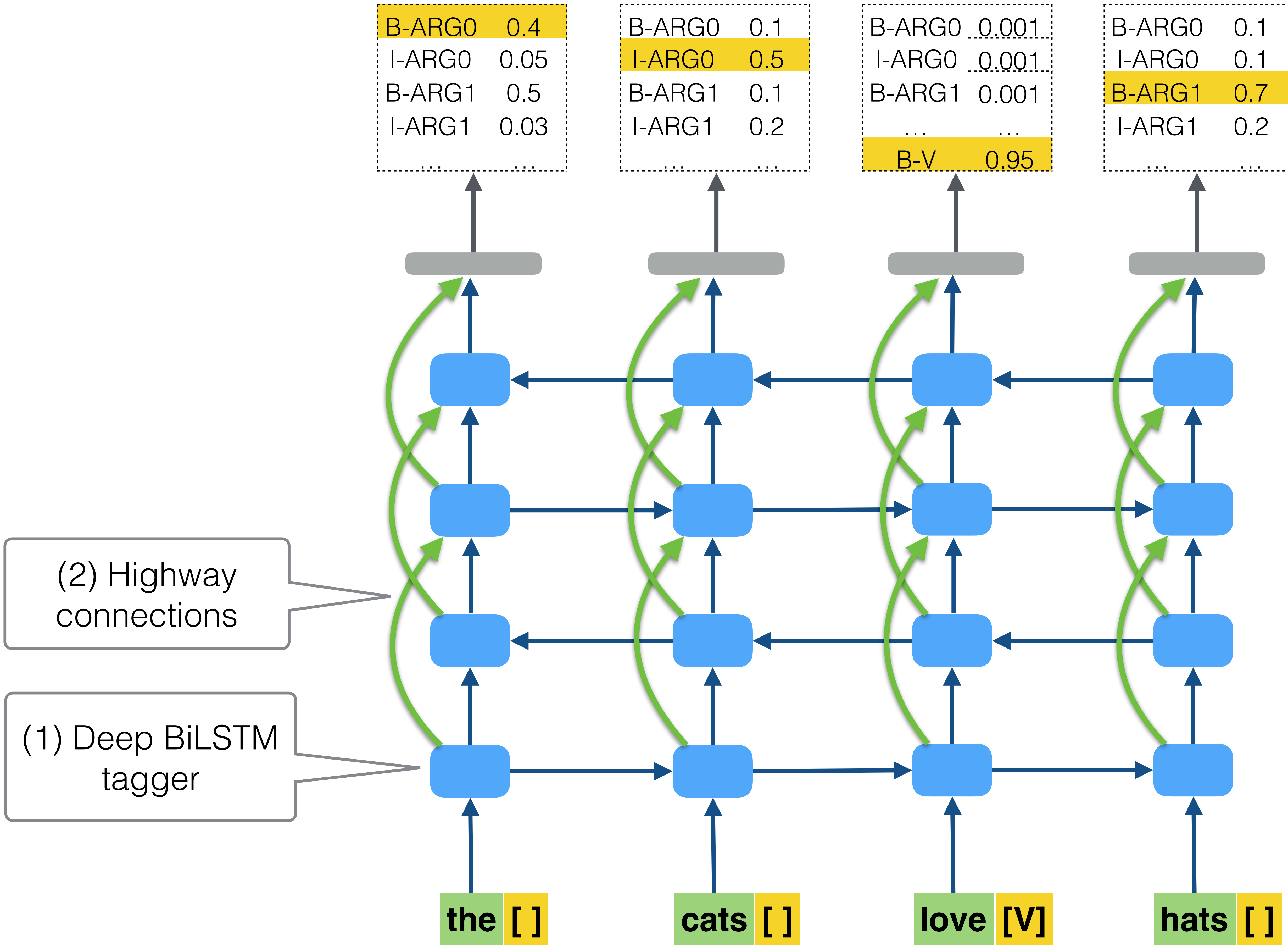
(**B**egin, **I**nside, **O**utside)

SRL as BIO Tagging Problem

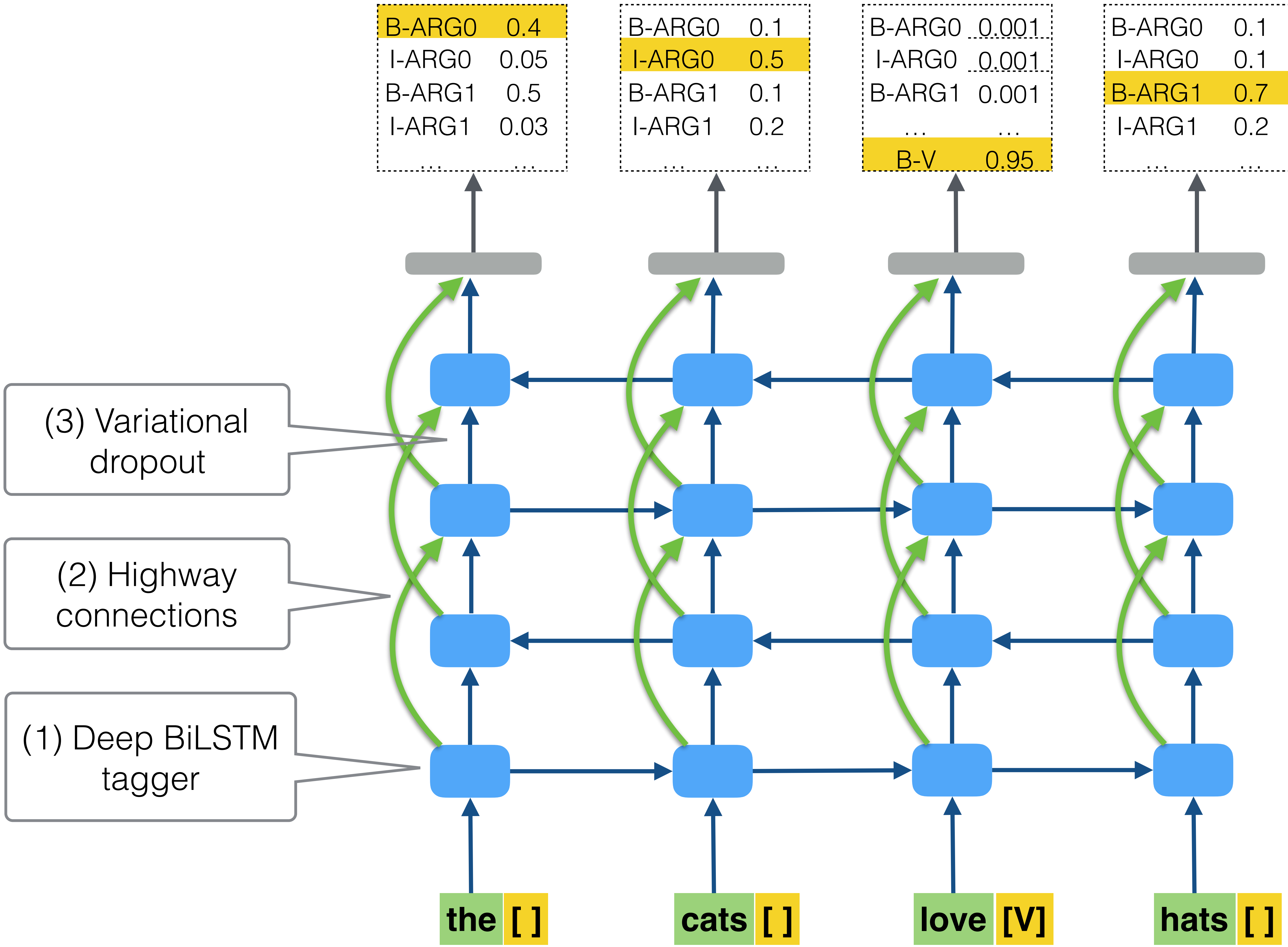


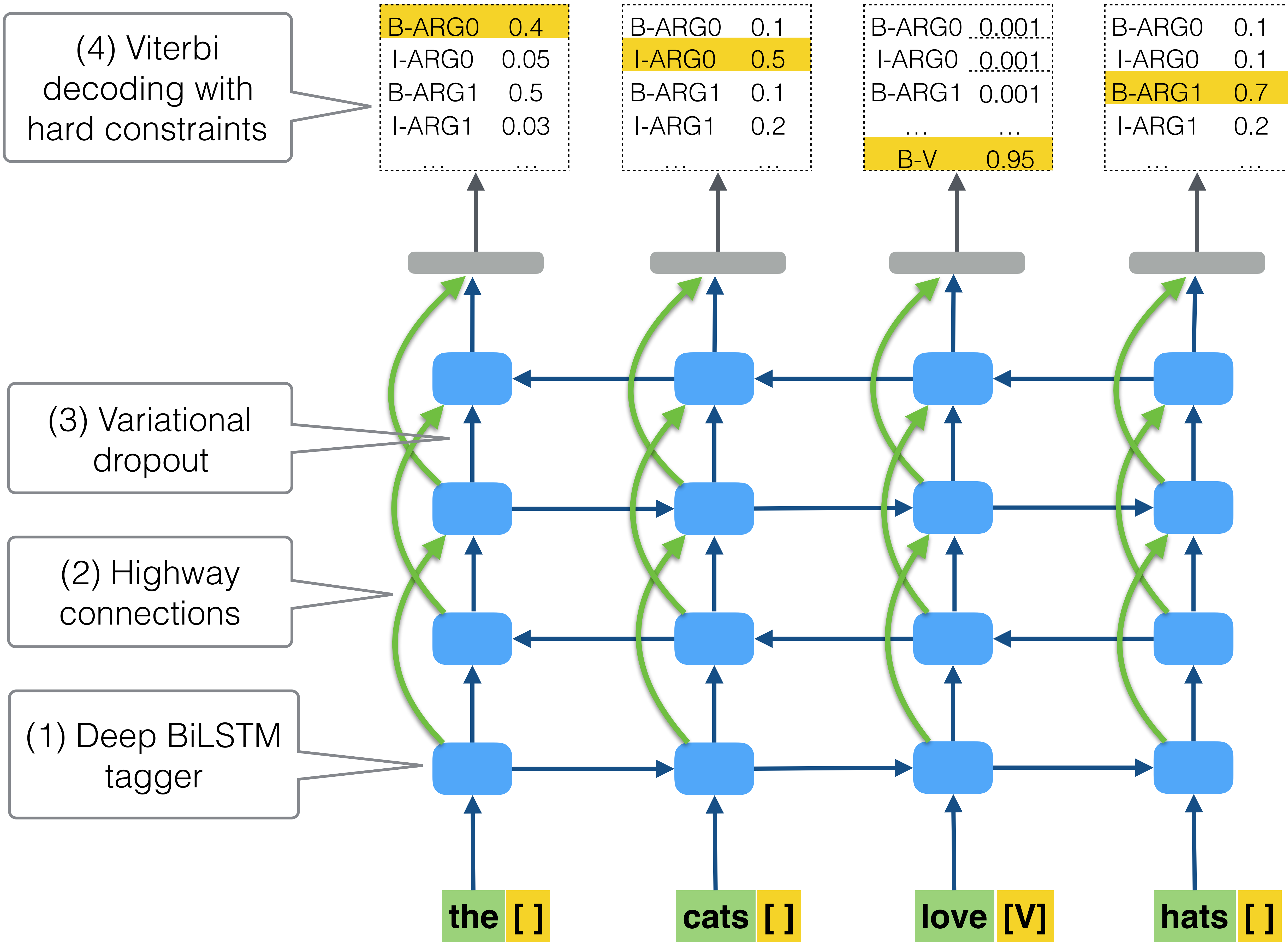

B-ARG0 0.4
I-ARG0 0.05
B-ARG1 0.5
I-ARG1 0.03
... ...
B-ARG0 0.1
I-ARG0 0.5
B-ARG1 0.1
I-ARG1 0.2
... ...
B-ARG0 0.001
I-ARG0 0.001
B-ARG1 0.001
... ...
B-V 0.95
B-ARG0 0.1
I-ARG0 0.1
B-ARG1 0.7
I-ARG1 0.2
... ...
the [ ]
cats [ ]
love [V]
hats [ ]


B-ARG0 0.4
I-ARG0 0.05
B-ARG1 0.5
I-ARG1 0.03
... ...
B-ARG0 0.1
I-ARG0 0.5
B-ARG1 0.1
I-ARG1 0.2
... ...
B-ARG0 0.001
I-ARG0 0.001
B-ARG1 0.001
... ...
B-V 0.95
B-ARG0 0.1
I-ARG0 0.1
B-ARG1 0.7
I-ARG1 0.2
... ...
(1) Deep BiLSTM tagger
the [ ]
cats [ ]
love [V]
hats [ ]



B-ARG0 0.4
I-ARG0 0.05
B-ARG1 0.5
I-ARG1 0.03
...
B-ARG0 0.1
I-ARG0 0.5
B-ARG1 0.1
I-ARG1 0.2
...
B-ARG0 0.001
I-ARG0 0.001
B-ARG1 0.001
...
B-V 0.95
B-ARG0 0.1
I-ARG0 0.1
B-ARG1 0.7
I-ARG1 0.2
...
(2) Highway connections
(1) Deep BiLSTM tagger
the [ ]
cats [ ]
love [V]
hats [ ]

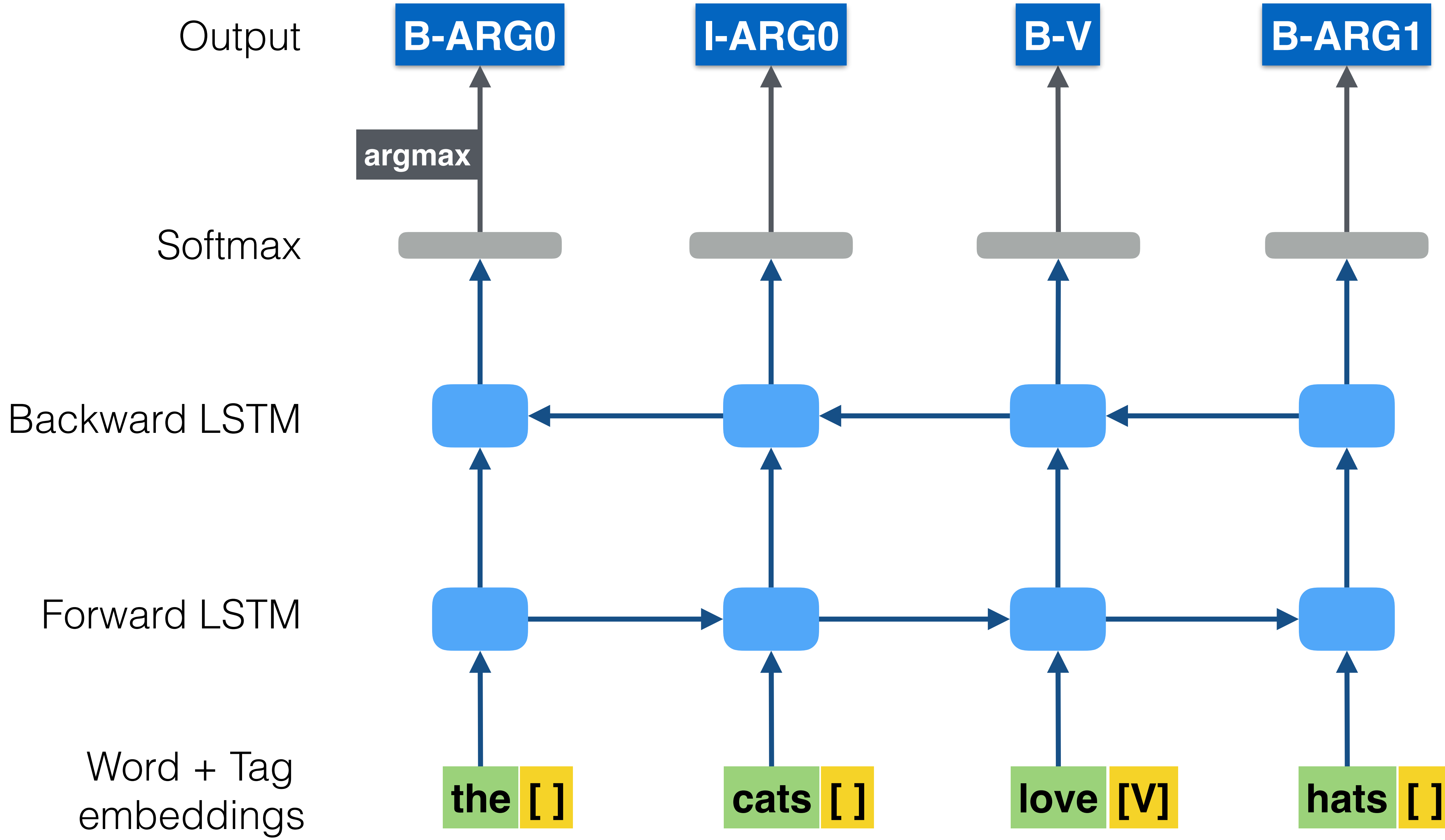

B-ARG0 0.4
I-ARG0 0.05
B-ARG1 0.5
I-ARG1 0.03
... ...
B-ARG0 0.1
I-ARG0 0.5
B-ARG1 0.1
I-ARG1 0.2
... ...
B-ARG0 0.001
I-ARG0 0.001
B-ARG1 0.001
... ...
B-V 0.95
B-ARG0 0.1
I-ARG0 0.1
B-ARG1 0.7
I-ARG1 0.2
... ...
(3) Variational dropout
(2) Highway connections
(1) Deep BiLSTM tagger
the [ ]
cats [ ]
love [V]
hats [ ]



(4) Viterbi decoding with hard constraints
(3) Variational dropout
(2) Highway connections
(1) Deep BiLSTM tagger
B-ARG0 0.4
I-ARG0 0.05
B-ARG1 0.5
I-ARG1 0.03
B-ARG0 0.1
I-ARG0 0.5
B-ARG1 0.1
I-ARG1 0.2
B-ARG0 0.001
I-ARG0 0.001
B-ARG1 0.001
B-V 0.95
B-ARG0 0.1
I-ARG0 0.1
B-ARG1 0.7
I-ARG1 0.2
the [ ]
cats [ ]
love [V]
hats [ ]

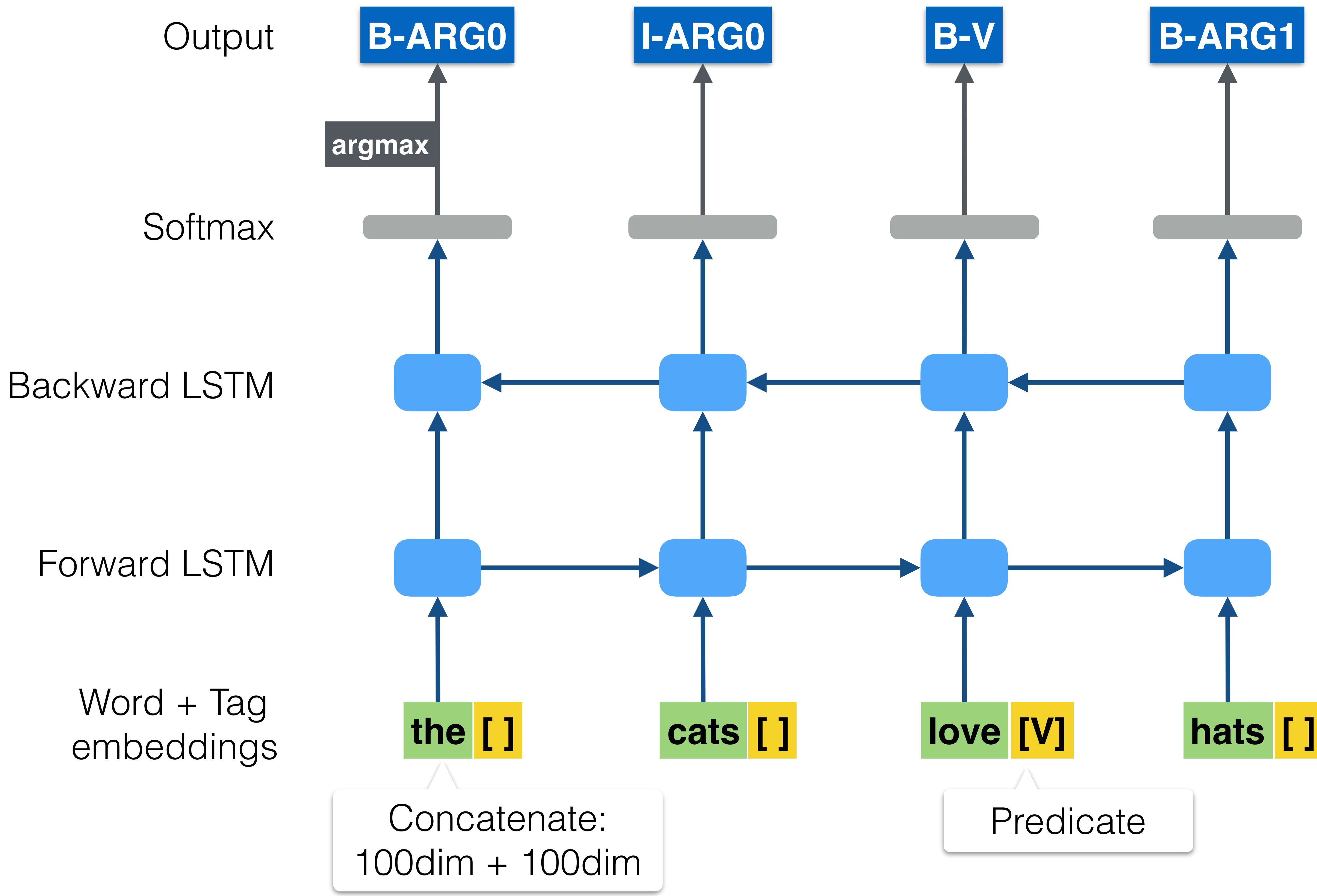
Deep BiLSTM Tagger

Highway Connections | Variational Dropout | Viterbi Decoding w\ Hard Constraints



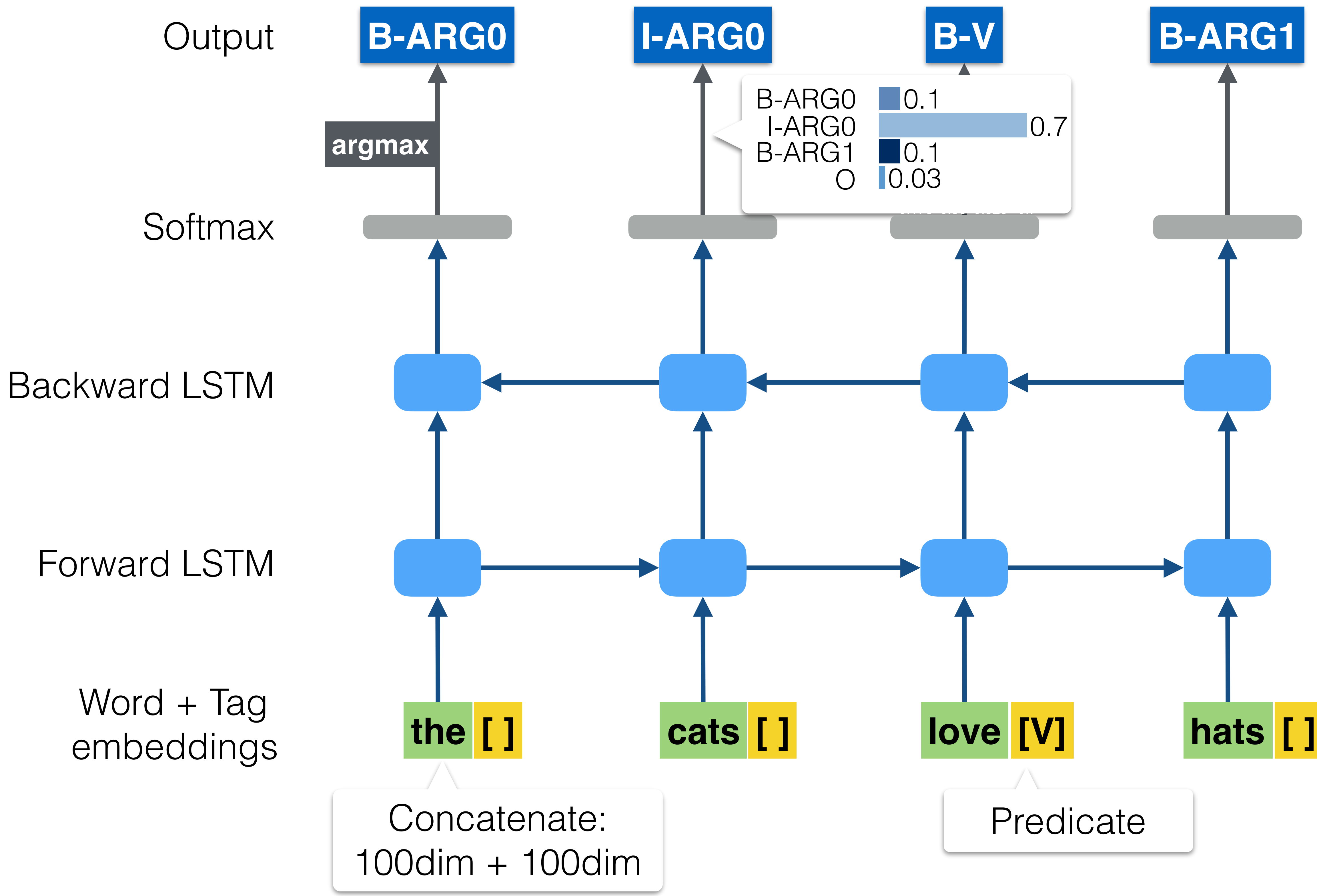
Deep BiLSTM Tagger

Highway Connections
Variational Dropout
Viterbi Decoding w\ Hard Constraints



Deep BiLSTM Tagger

Highway Connections | Variational Dropout | Viterbi Decoding w\ Hard Constraints



Highway Connections

Trend: Deeper models for higher accuracy

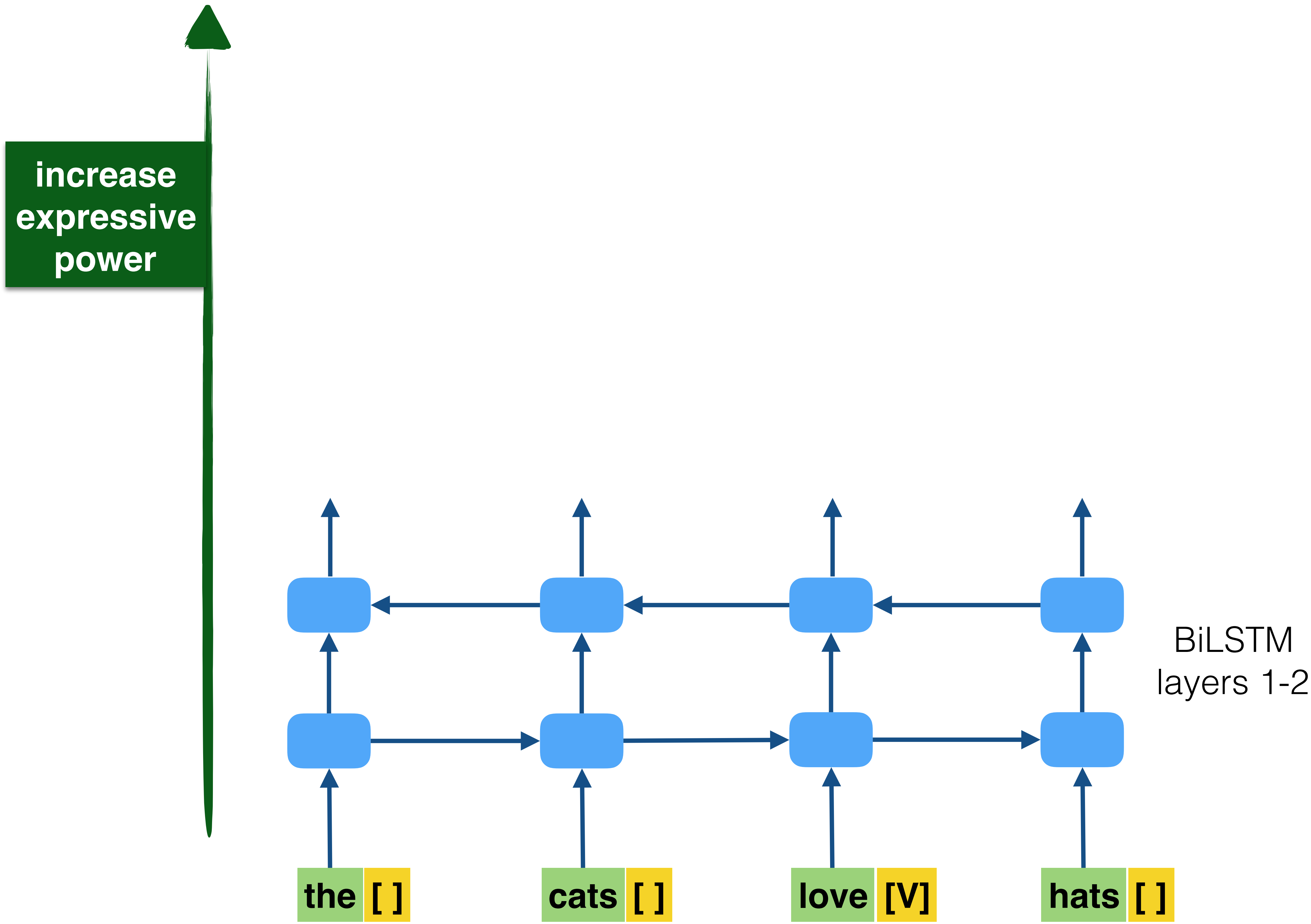
Grammar as a Foreign Language (Vinyals et al., 2014): **3** layers

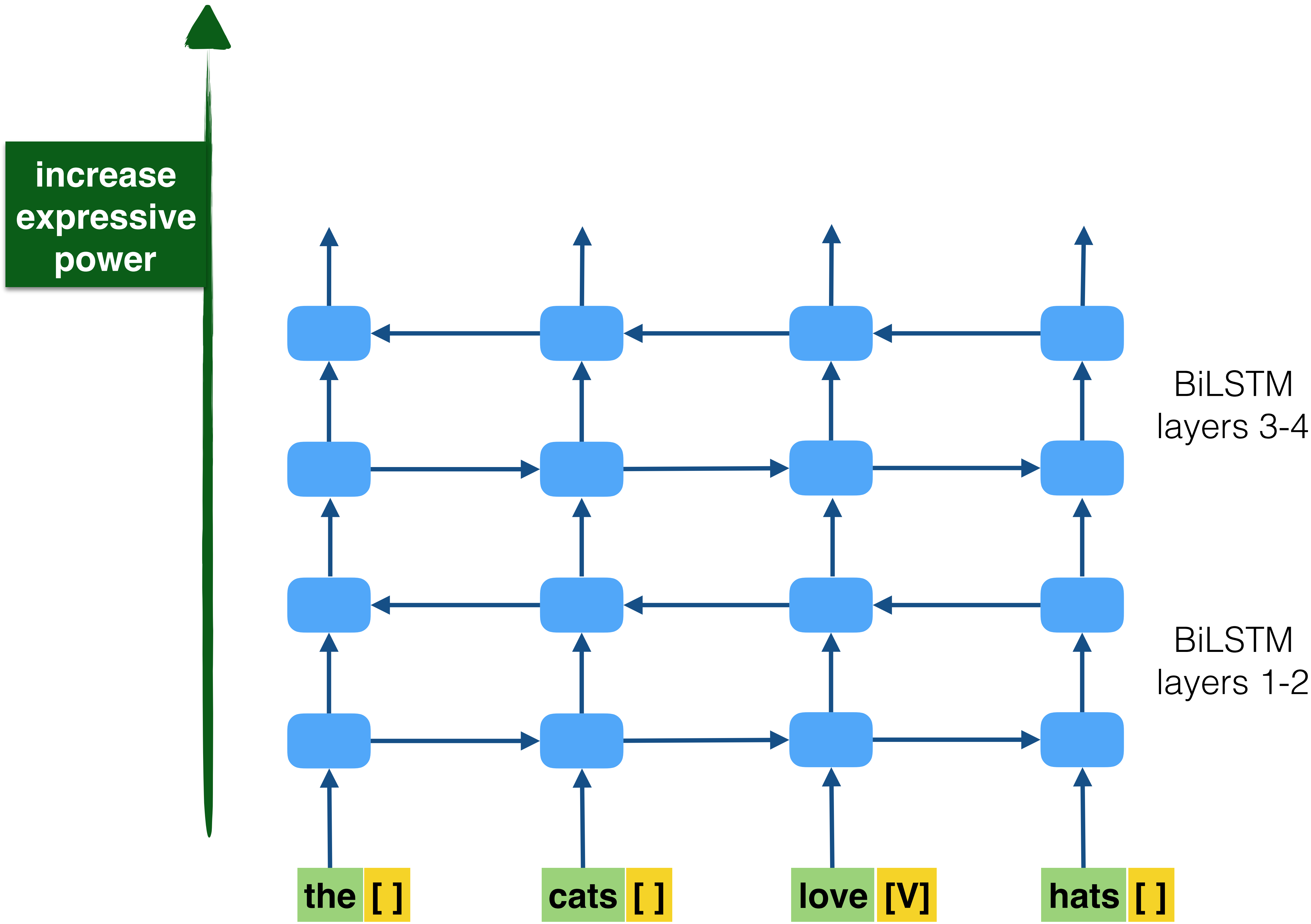
End-to-end Semantic Role Labeling (Zhou and Xu, 2015): **8** layers

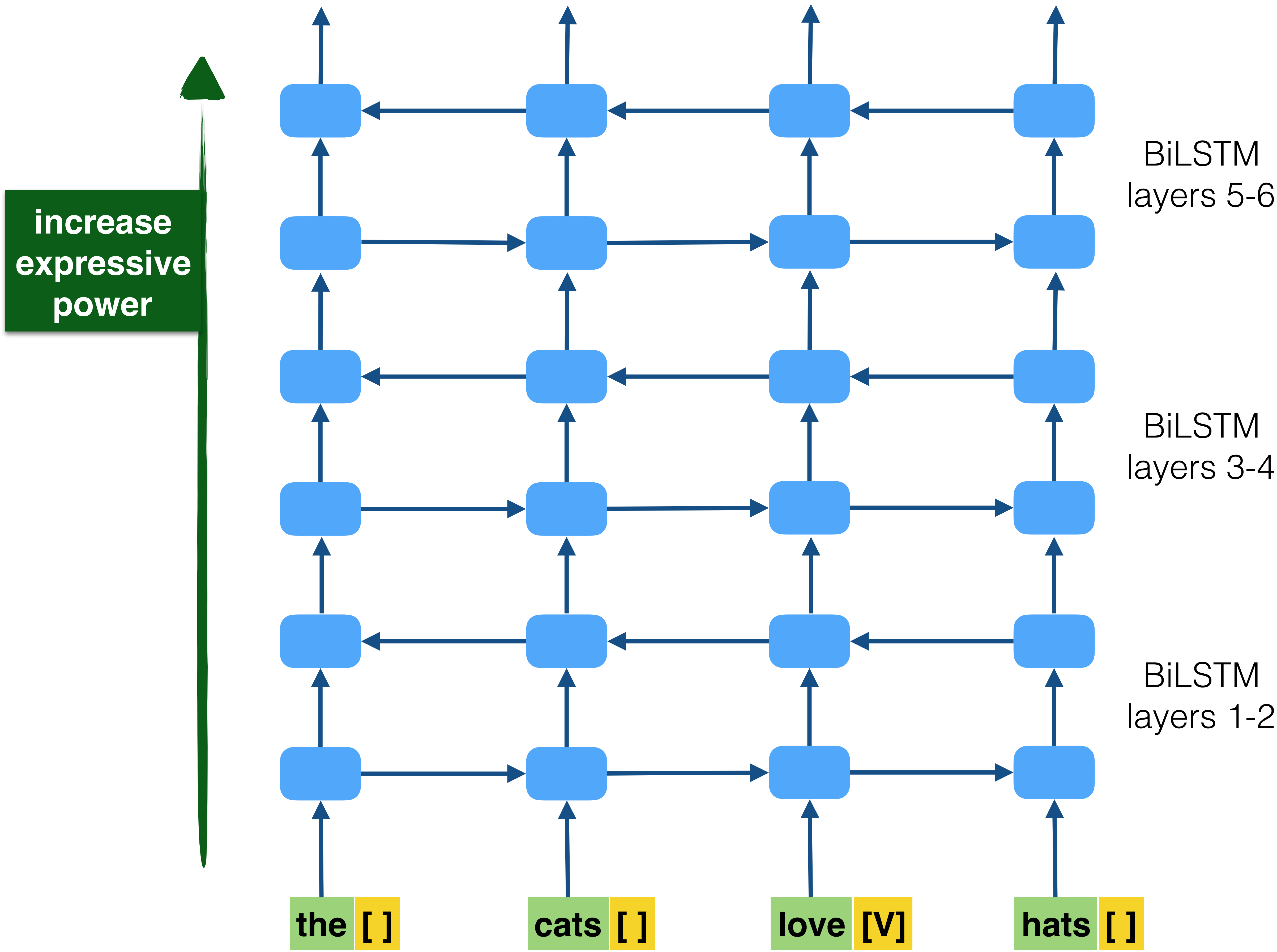
Google's Neural Machine Translation (GNMT, Wu et al., 2016): **8** layers

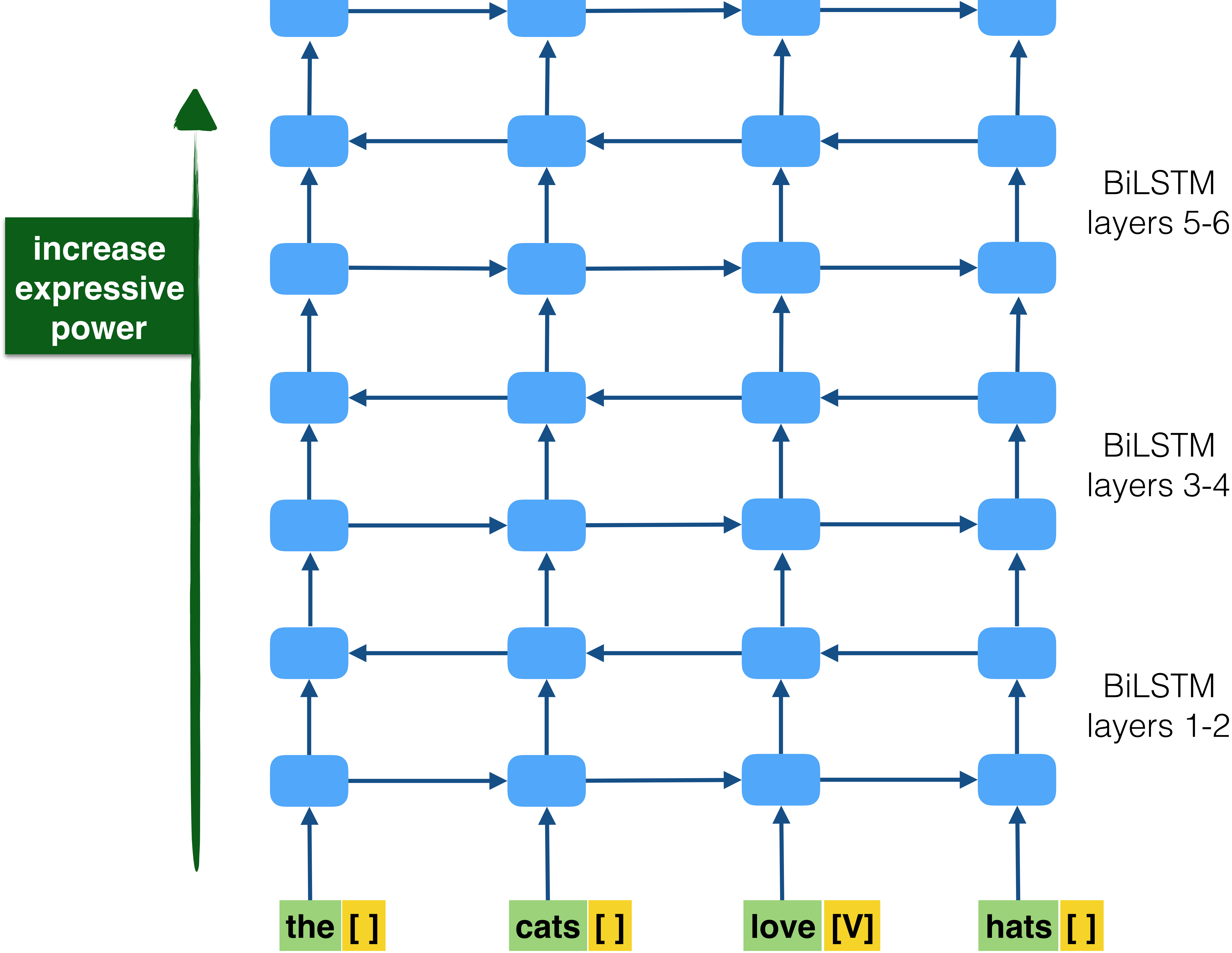
this work: **8** layers

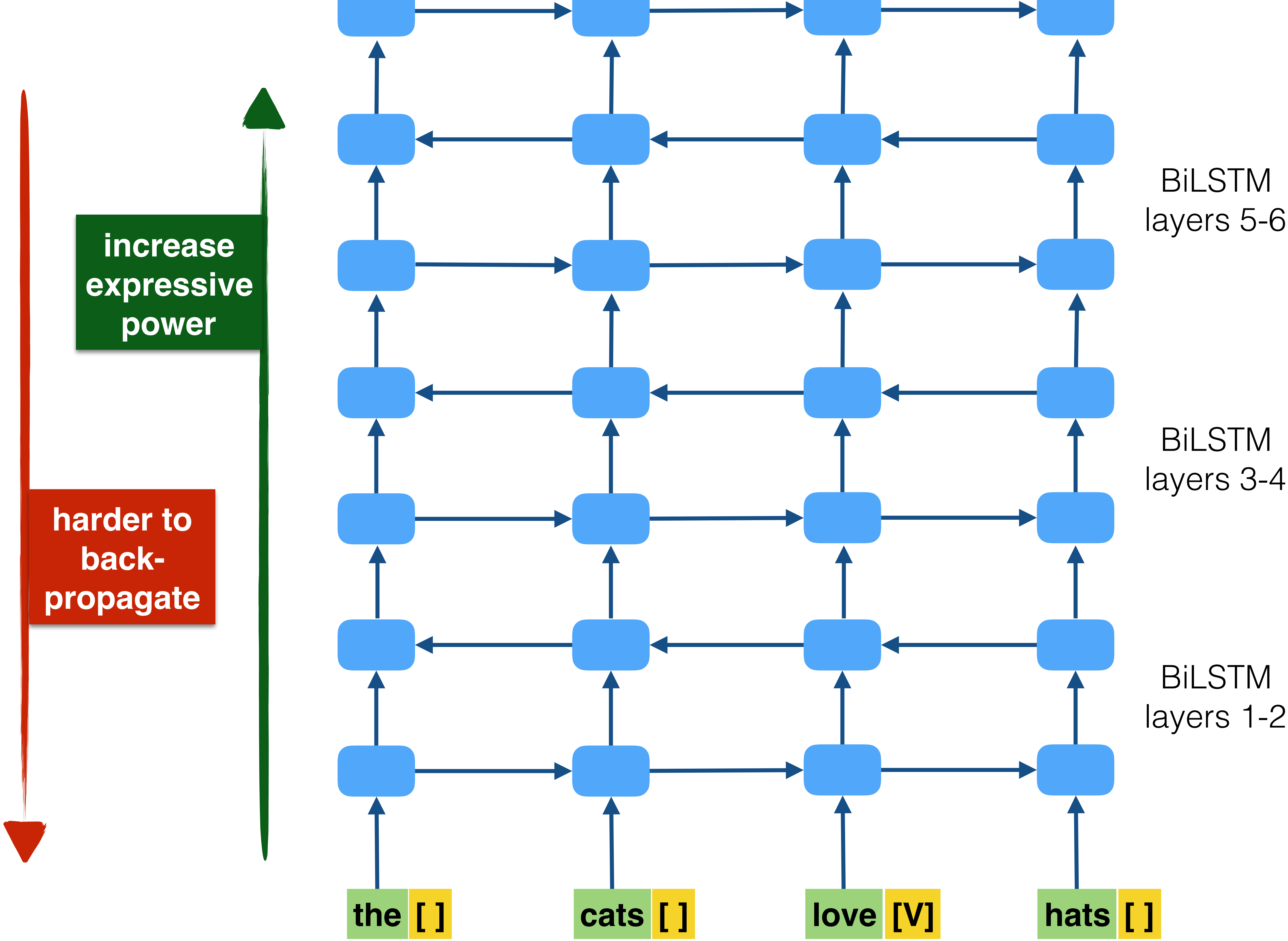
Deep Residual Learning for Image Recognition (He et al, 2016): **152** layers


increase expressive power
the [ ]
cats [ ]
love [V]
hats [ ]
BiLSTM layers 1-2


increase expressive power
BiLSTM layers 3-4
BiLSTM layers 1-2
the [ ]
cats [ ]
love [V]
hats [ ]


increase expressive power
BiLSTM layers 5-6
BiLSTM layers 3-4
BiLSTM layers 1-2
the [ ]
cats [ ]
love [V]
hats [ ]


increase expressive power
BiLSTM layers 5-6
BiLSTM layers 3-4
BiLSTM layers 1-2
the [ ]
cats [ ]
love [V]
hats [ ]



increase expressive power
harder to back-propagate
BiLSTM layers 5-6
BiLSTM layers 3-4
BiLSTM layers 1-2
the [ ]
cats [ ]
love [V]
hats [ ]

Highway Connections

Grammar as a Foreign Language (Vinyals et al., 2014): **3** layers

End-to-end Semantic Role Labeling (Zhou and Xu, 2015): **8** layers

Google's Neural Machine Translation (GNMT, Wu et al., 2016): **8** layers

this work: **8** layers

Deep Residual Learning for Image Recognition (He et al, 2016): **152** layers

Highway Connections

Grammar as a Foreign Language (Vinyals et al., 2014): **3** layers

End-to-end Semantic Role Labeling (Zhou and Xu, 2015): **8** layers

Google's Neural Machine Translation (GNMT, Wu et al., 2016): **8** layers

this work: **8** layers

Deep Residual Learning for Image Recognition (He et al, 2016): **152** layers

use different learning rates for different layers (harder to reimplement)

Highway Connections

Grammar as a Foreign Language (Vinyals et al., 2014): **3** layers

End-to-end Semantic Role Labeling (Zhou and Xu, 2015): **8** layers

Google's Neural Machine Translation (GNMT, Wu et al., 2016): **8** layers

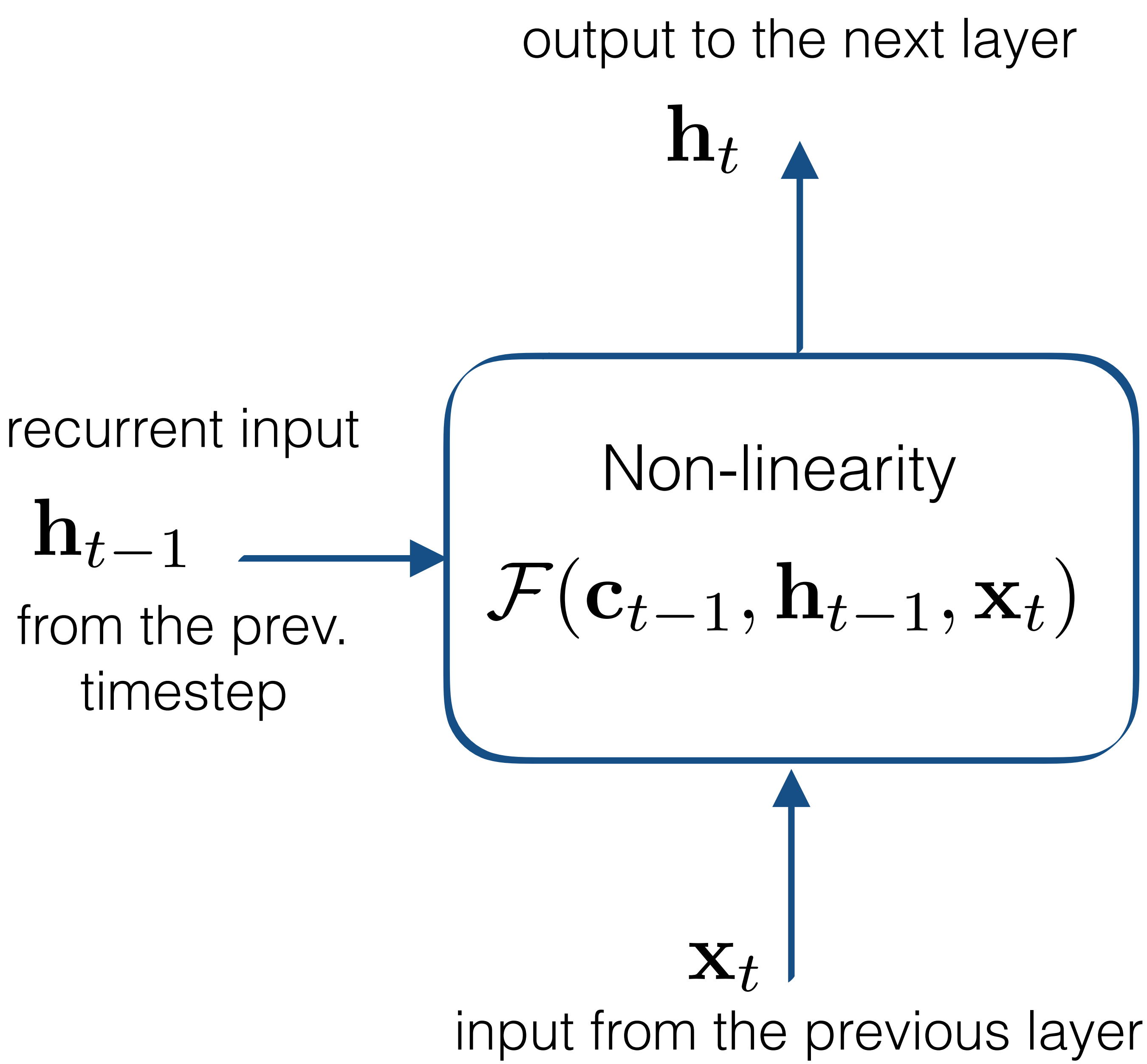
this work: **8** layers

Deep Residual Learning for Image Recognition (He et al, 2016): **152** layers

use different learning rates for different layers (harder to reimplement)

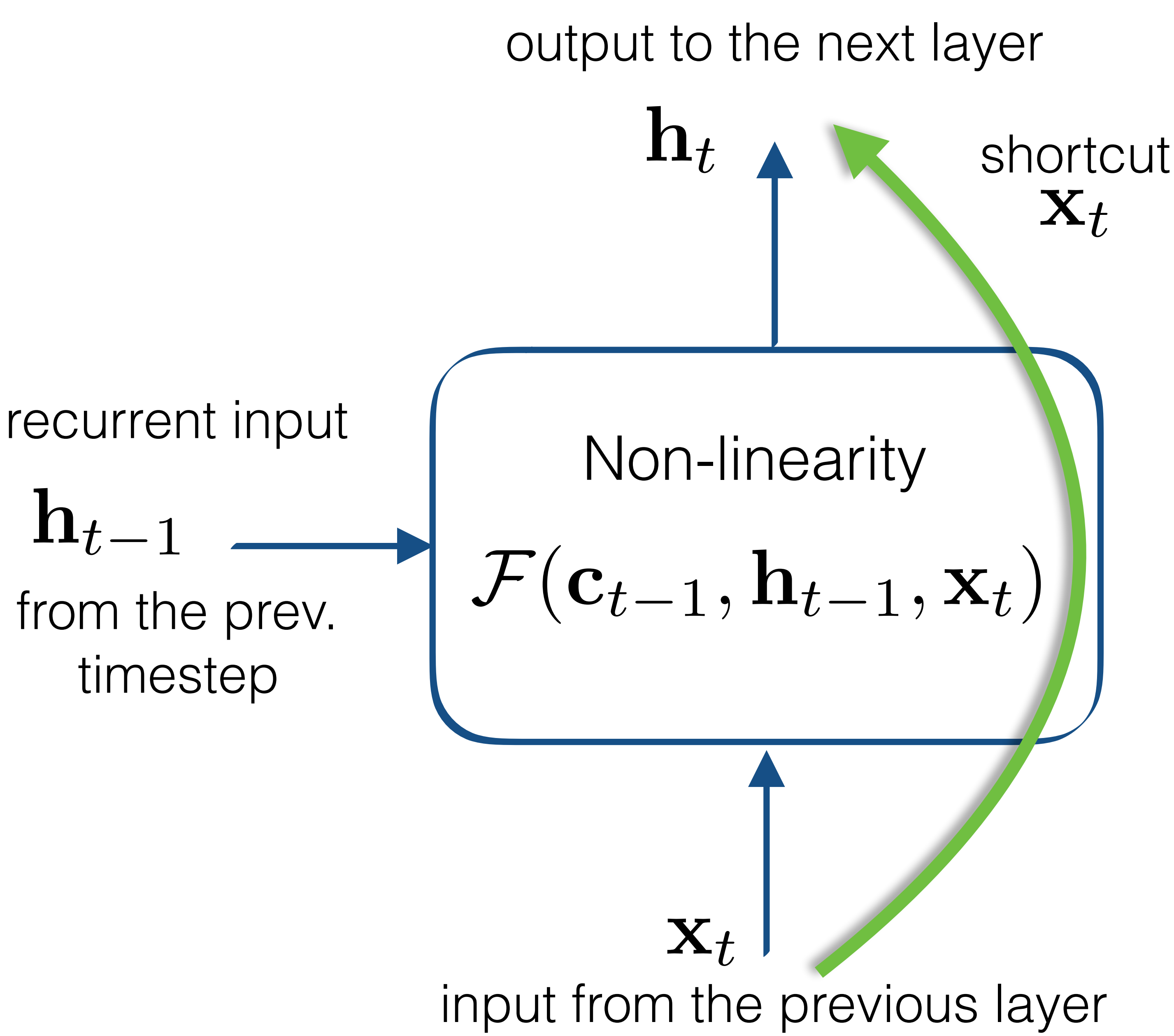
use shortcut connections between layers ("highway" or "residual")

Highway Connections



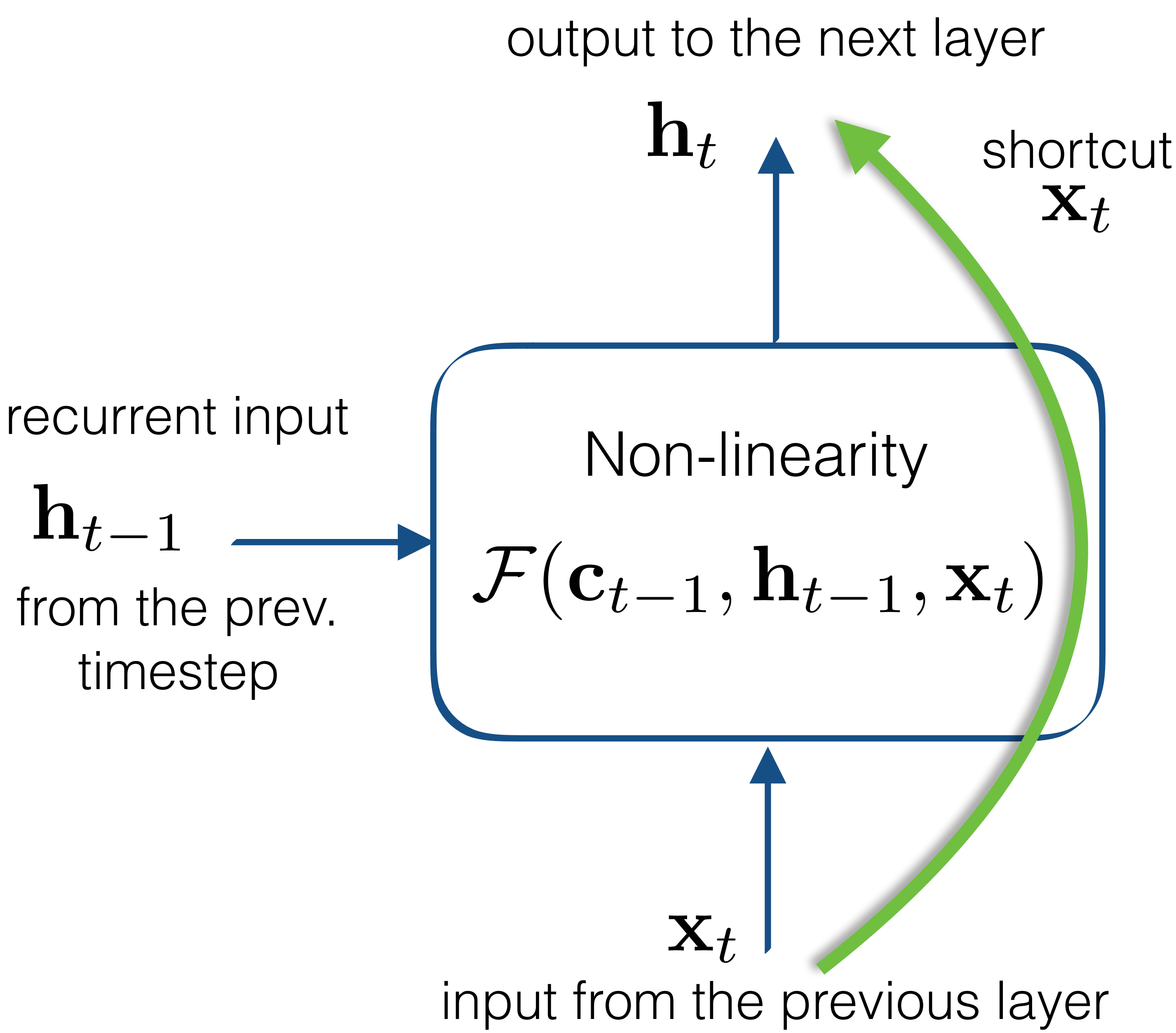
References:
Deep Residual Networks, Kaiming He, ICML 2016 Tutorial
Training Very Deep Networks, Srivastava et al., 2015

Highway Connections



References:
Deep Residual Networks, Kaiming He, ICML 2016 Tutorial
Training Very Deep Networks, Srivastava et al., 2015

Highway Connections

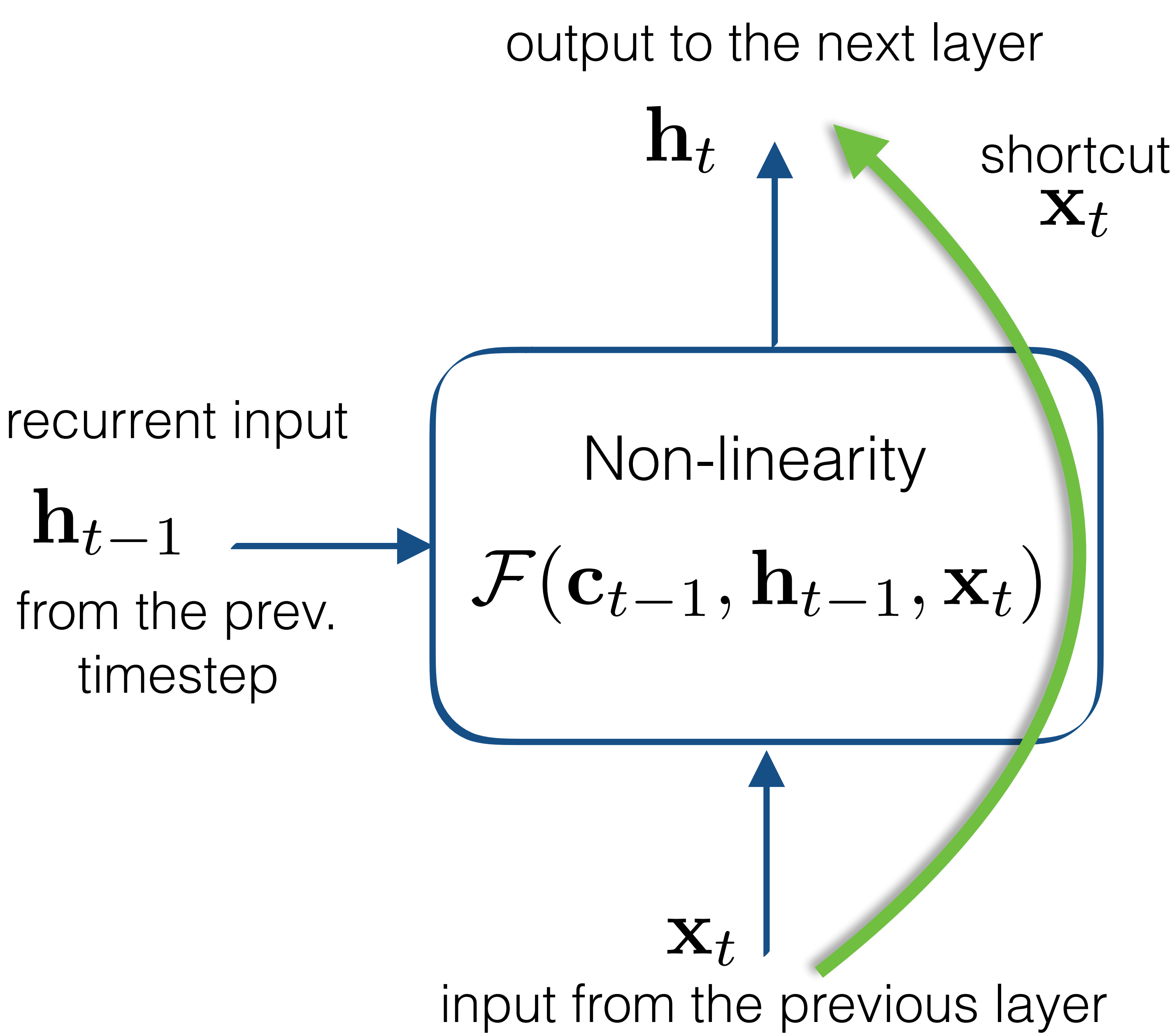


new output:

residual net $\mathbf{h}_t + \mathbf{x}_t$

References:
Deep Residual Networks, Kaiming He, ICML 2016 Tutorial
Training Very Deep Networks, Srivastava et al., 2015

Highway Connections



new output:

residual net $\mathbf{h}_t + \mathbf{x}_t$

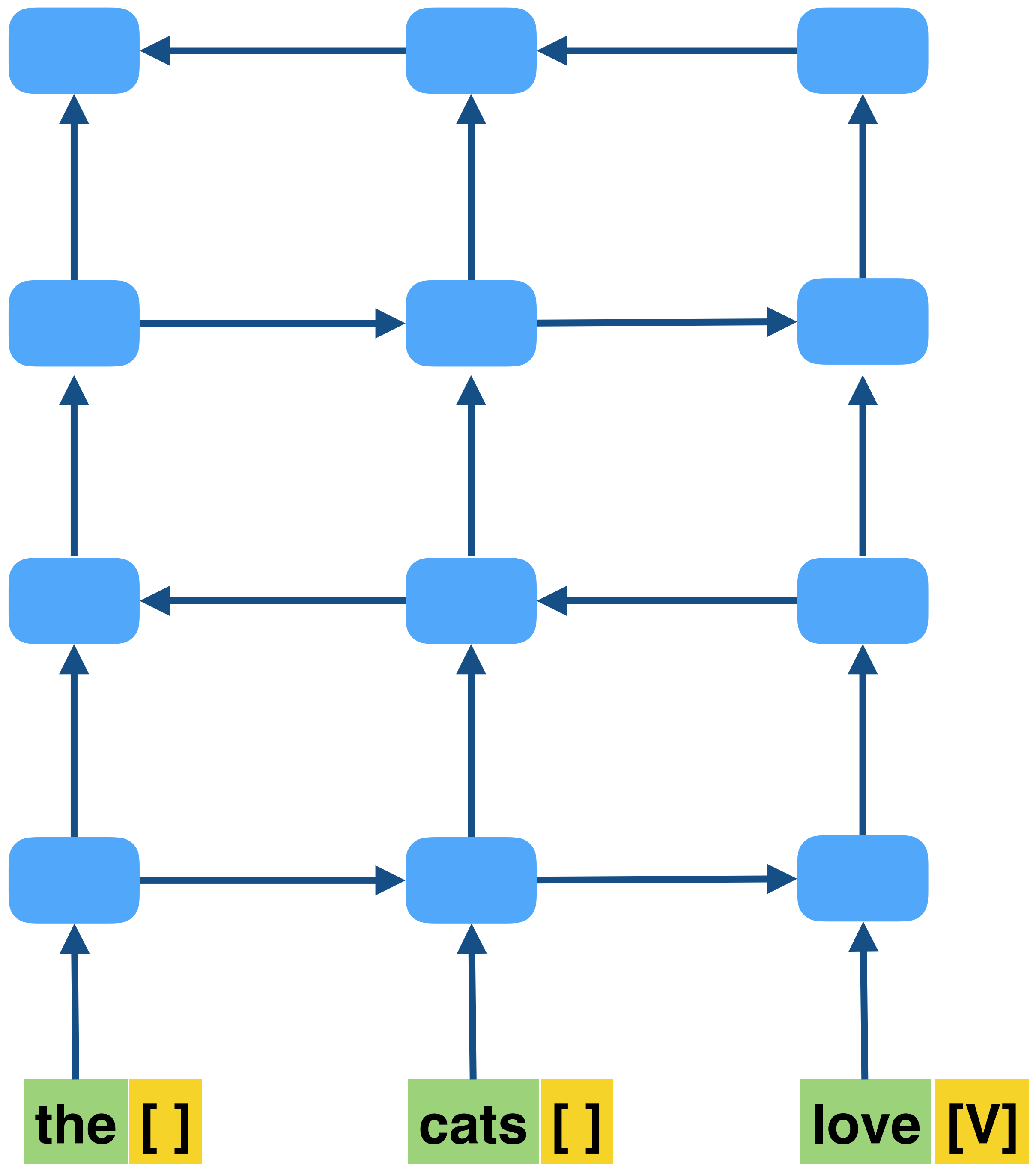
gated highway network:

$$\mathbf{r}_t \circ \mathbf{h}_t + (1 - \mathbf{r}_t) \circ \mathbf{x}_t$$

$$\mathbf{r}_t = \sigma(f(\mathbf{h}_{t-1}, \mathbf{x}_t))$$

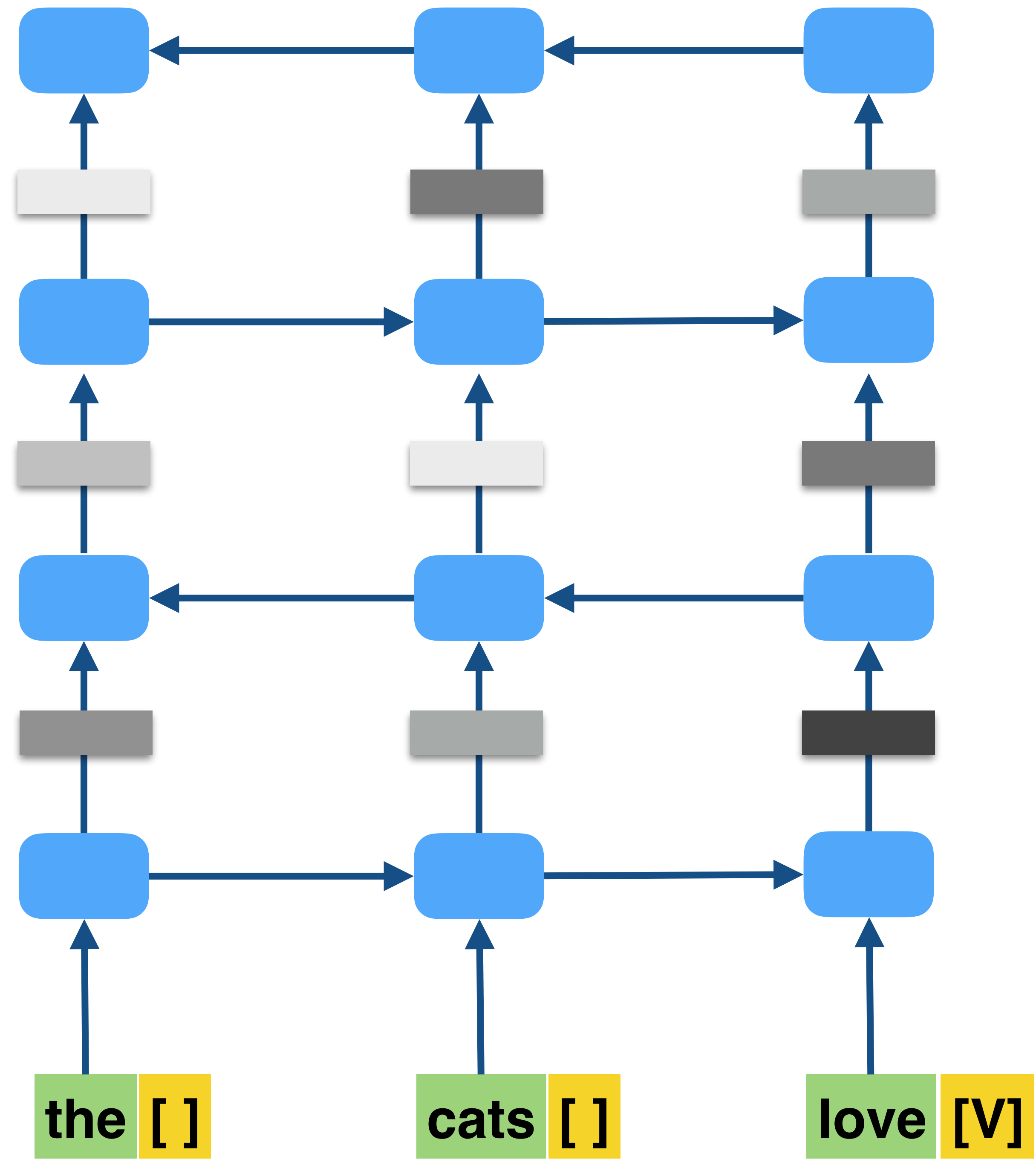
References:
Deep Residual Networks, Kaiming He, ICML 2016 Tutorial
Training Very Deep Networks, Srivastava et al., 2015

Variational Dropout



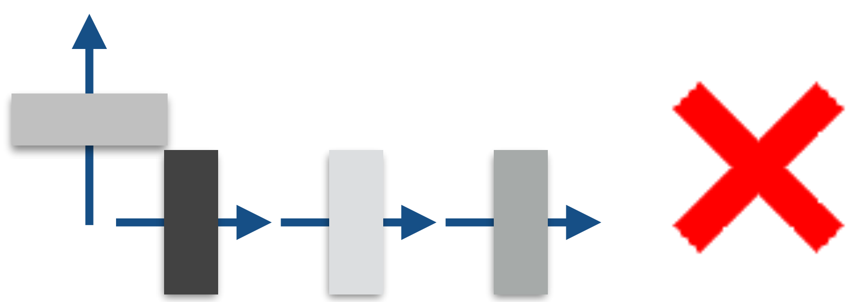
Variational Dropout

Traditionally, dropout masks are only applied to vertical connections.

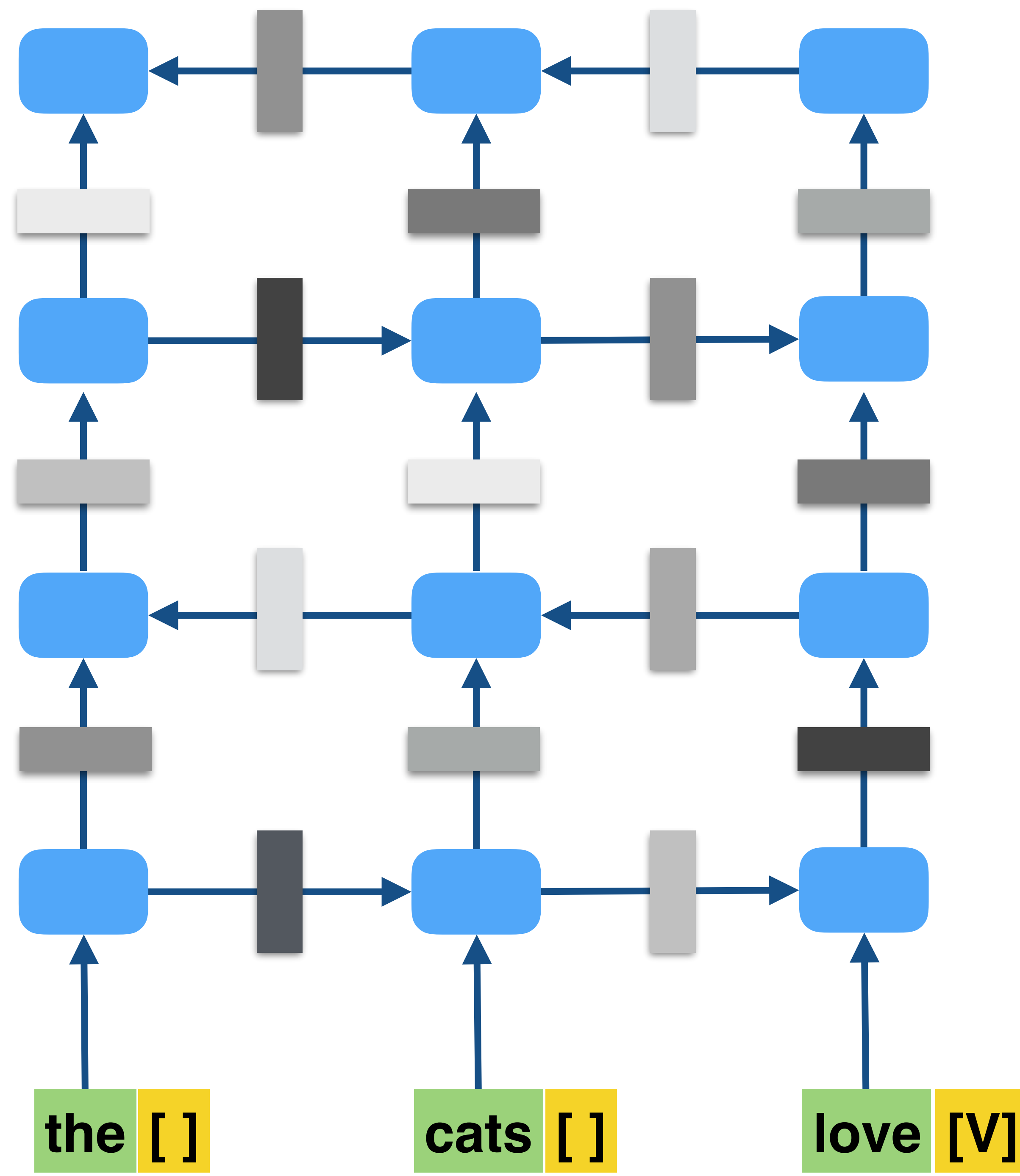


Variational Dropout

Traditionally, dropout masks are only applied to vertical connections.

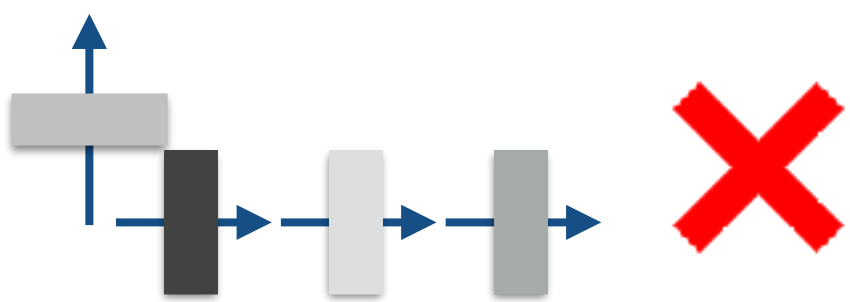


Applying dropout to recurrent connections causes too much noise amplification.

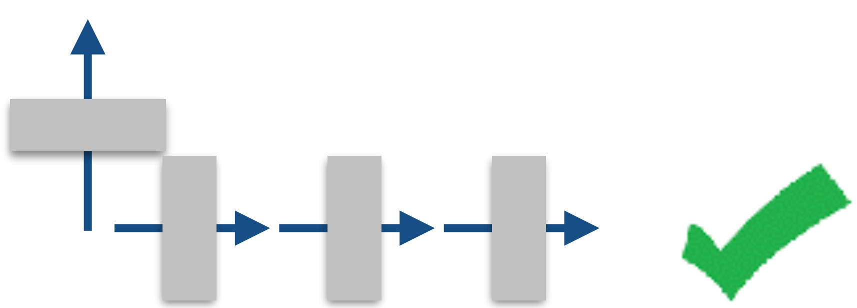


Variational Dropout

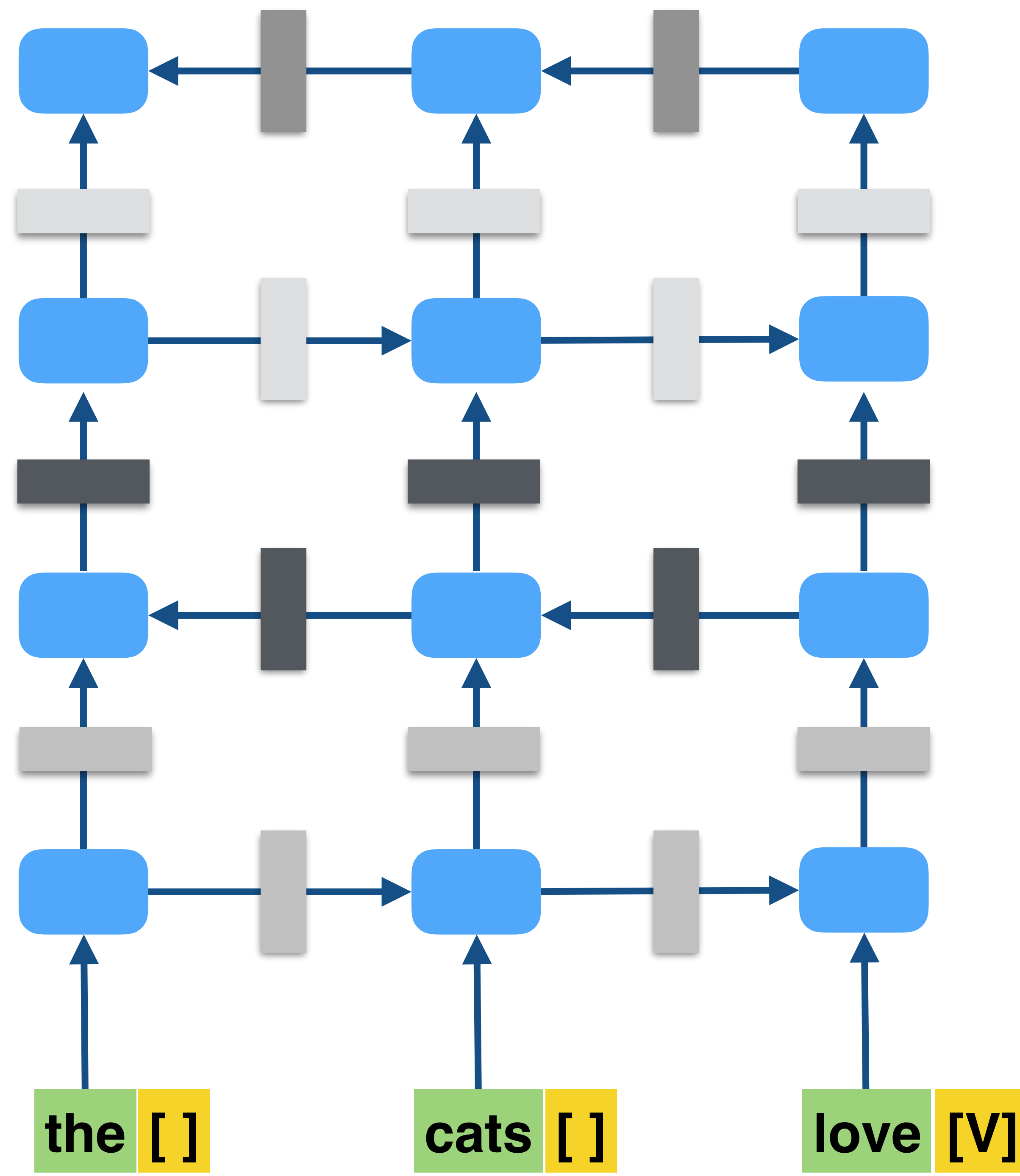
Traditionally, dropout masks are only applied to vertical connections.



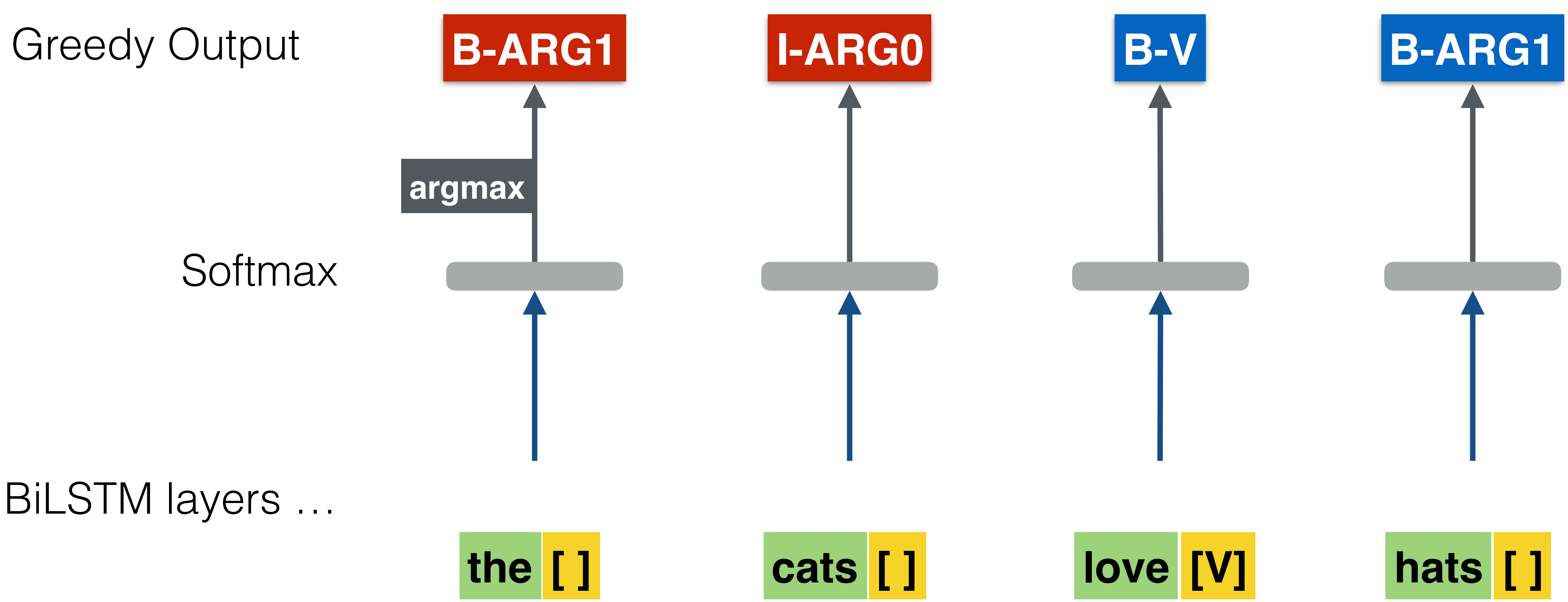
Applying dropout to recurrent connections causes too much noise amplification.



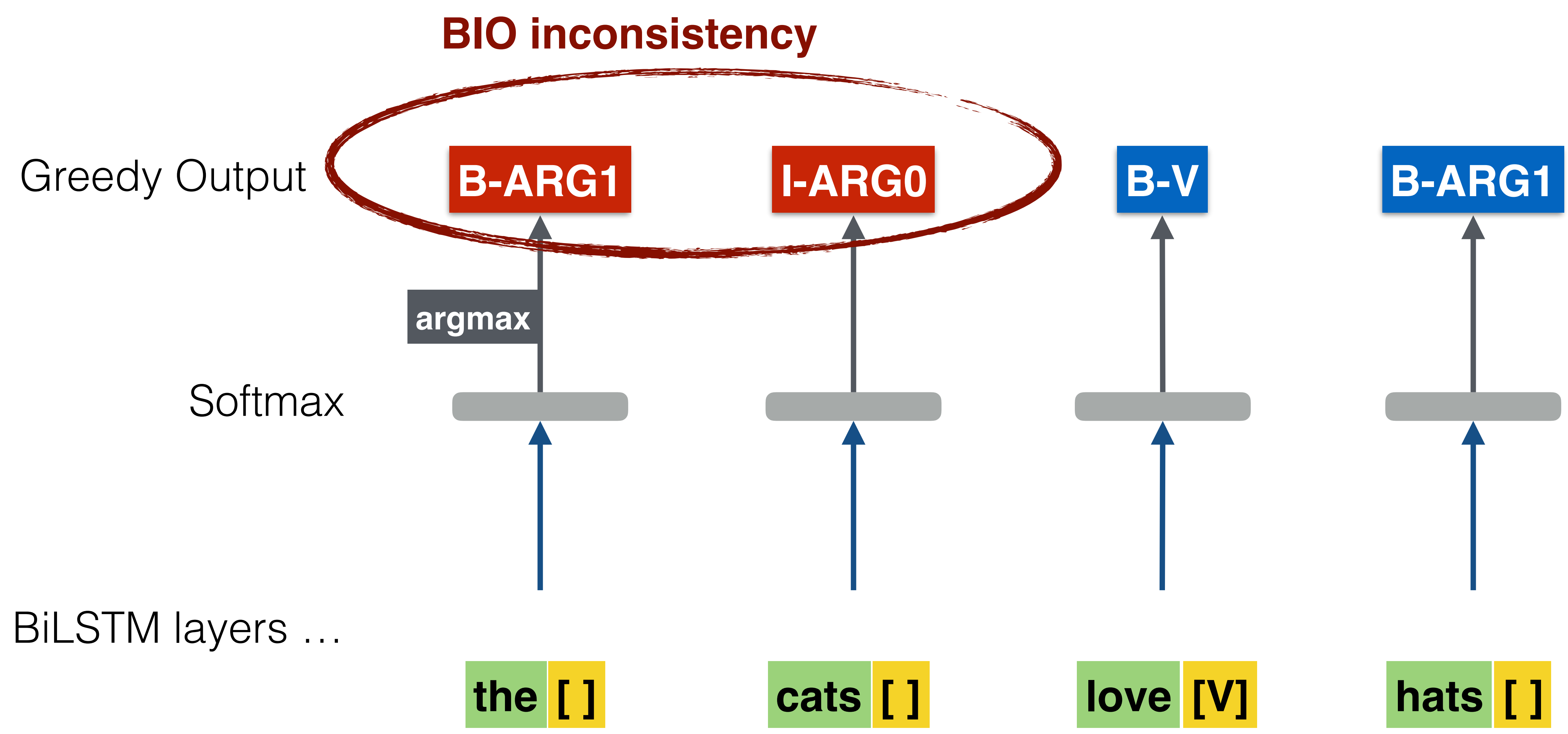
Variational dropout: Reuse the same dropout mask for each timestep.
Gal and Ghahramani, 2016



Viterbi Decoding w\ Hard Constraints



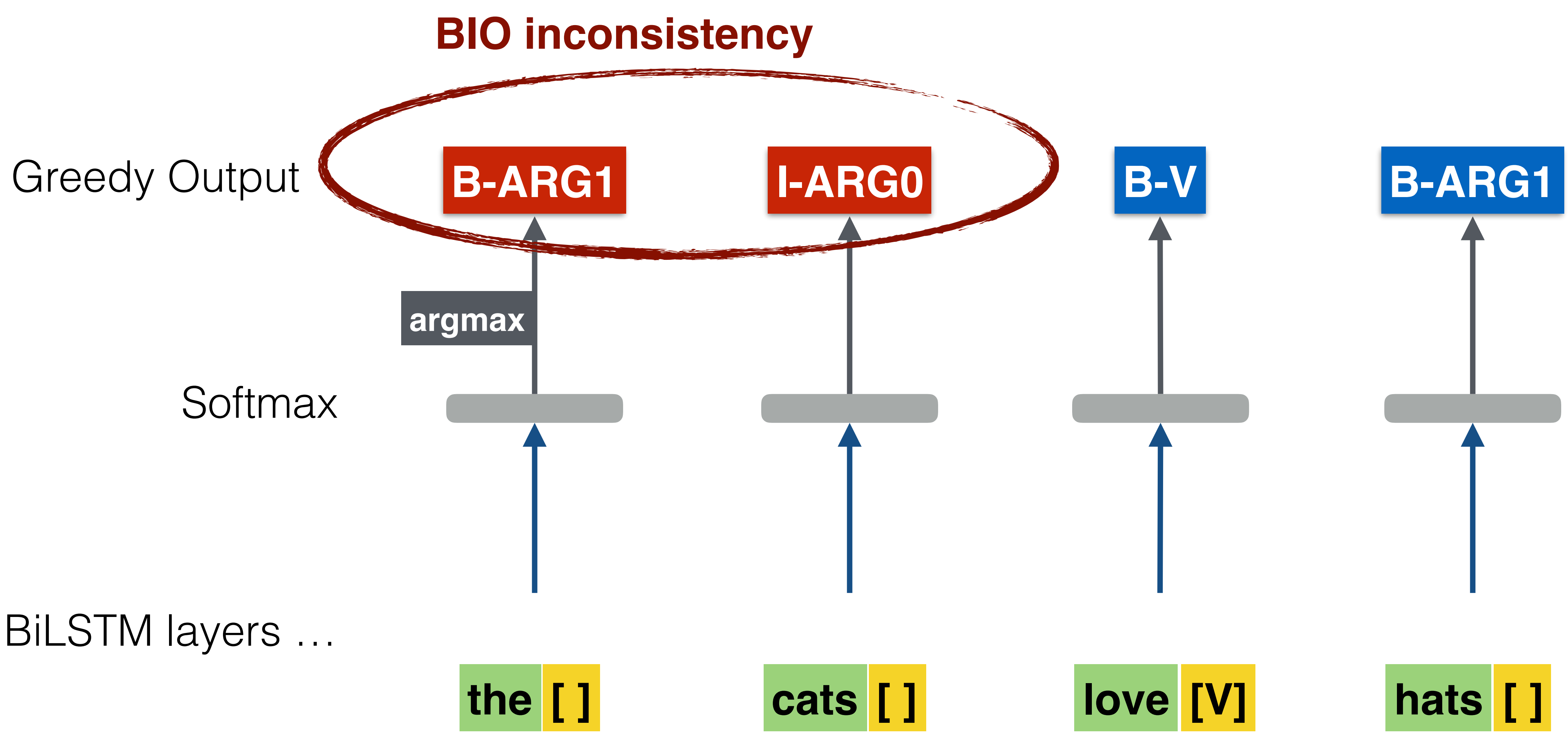
Viterbi Decoding w\ Hard Constraints



Viterbi Decoding w\ Hard Constraints

Heuristic transition scores

$$s(\text{B-ARG0} \rightarrow \text{I-ARG0}) = 0$$
$$s(\text{B-ARG1} \rightarrow \text{I-ARG0}) = -\inf$$
$$\ldots$$



Viterbi Decoding w\ Hard Constraints

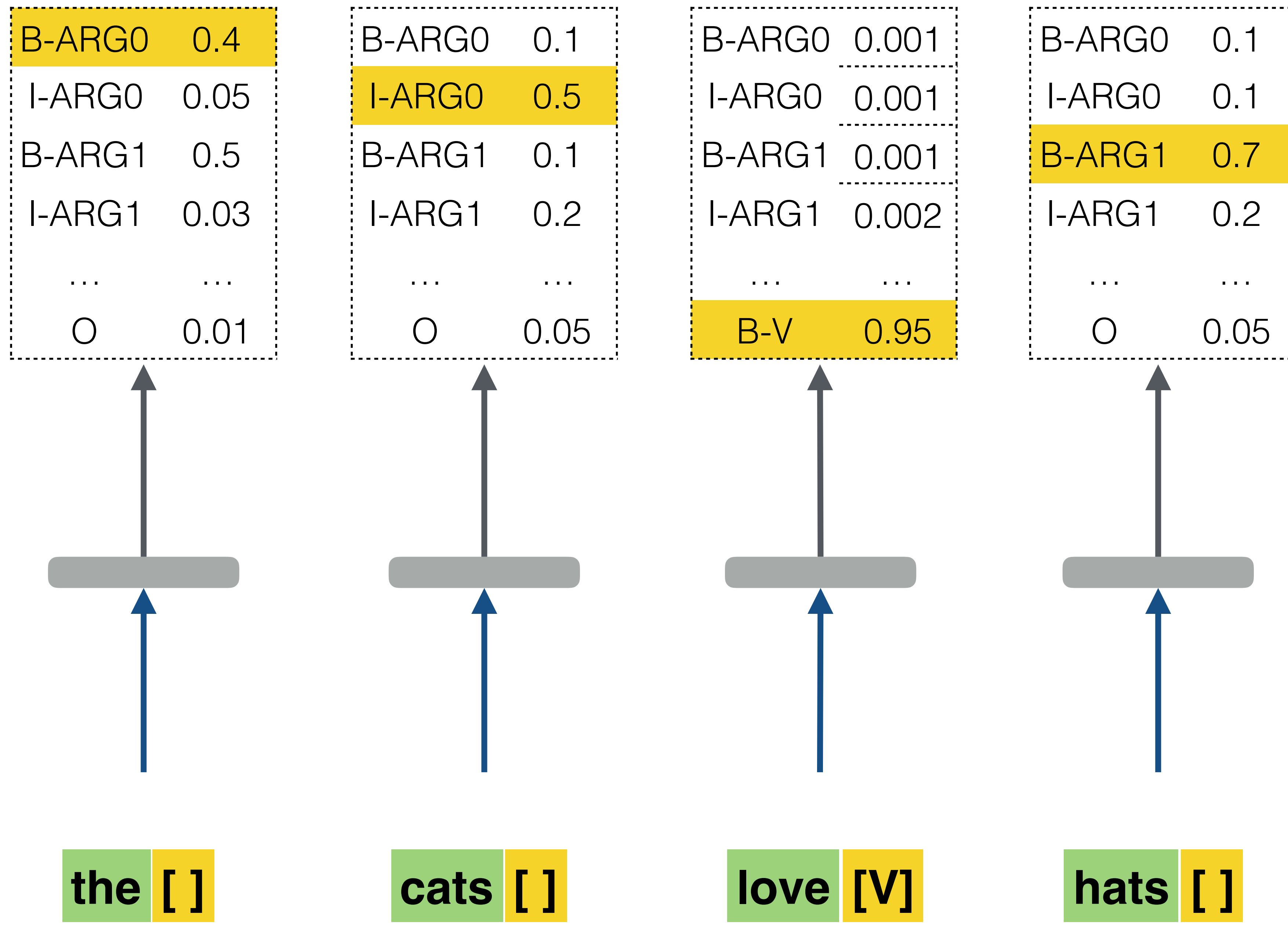
Heuristic transition scores

$$s(\text{B-ARG0} \rightarrow \text{I-ARG0}) = 0$$
$$s(\text{B-ARG1} \rightarrow \text{I-ARG0}) = -\inf$$
$$\ldots$$

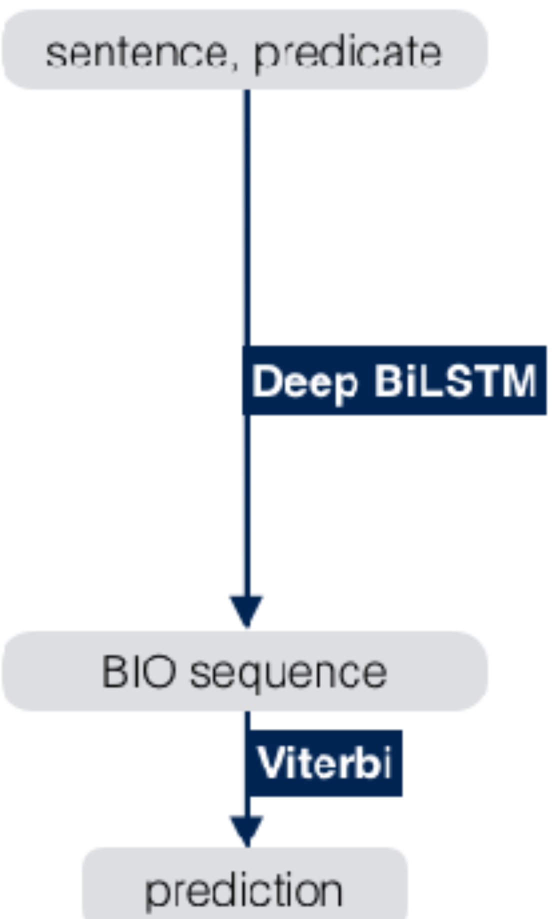
Viterbi decoding

Softmax

BiLSTM layers …



Other Implementation Details …

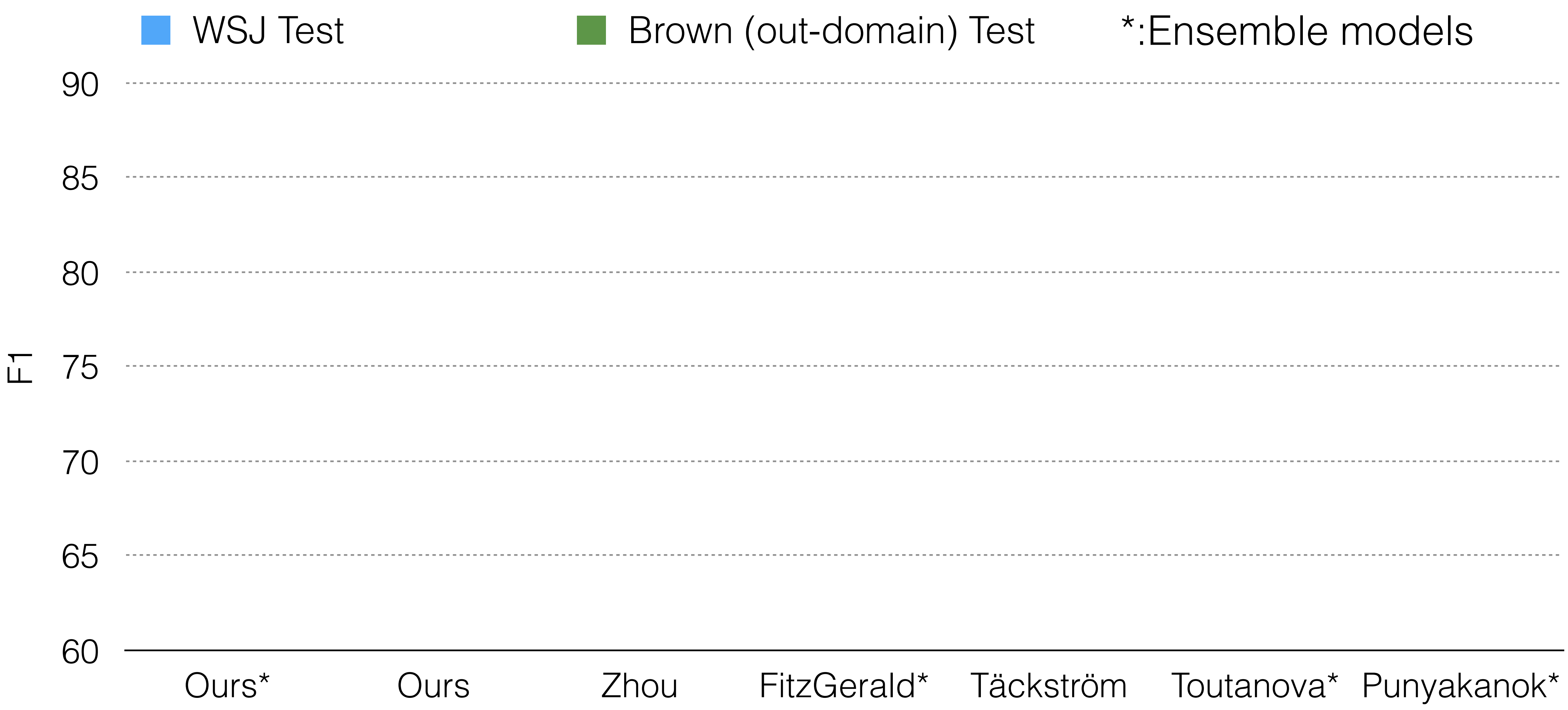


- 8 layer BiLSTMs with 300D hidden layers.
- 100D GloVe embeddings, updated during training.
- **Orthonormal initialization** for LSTM weight matrices (Saxe et al., 2013)
- 5 model ensemble with **product-of-experts** (Hinton 2002)
- Trained for 500 epochs.

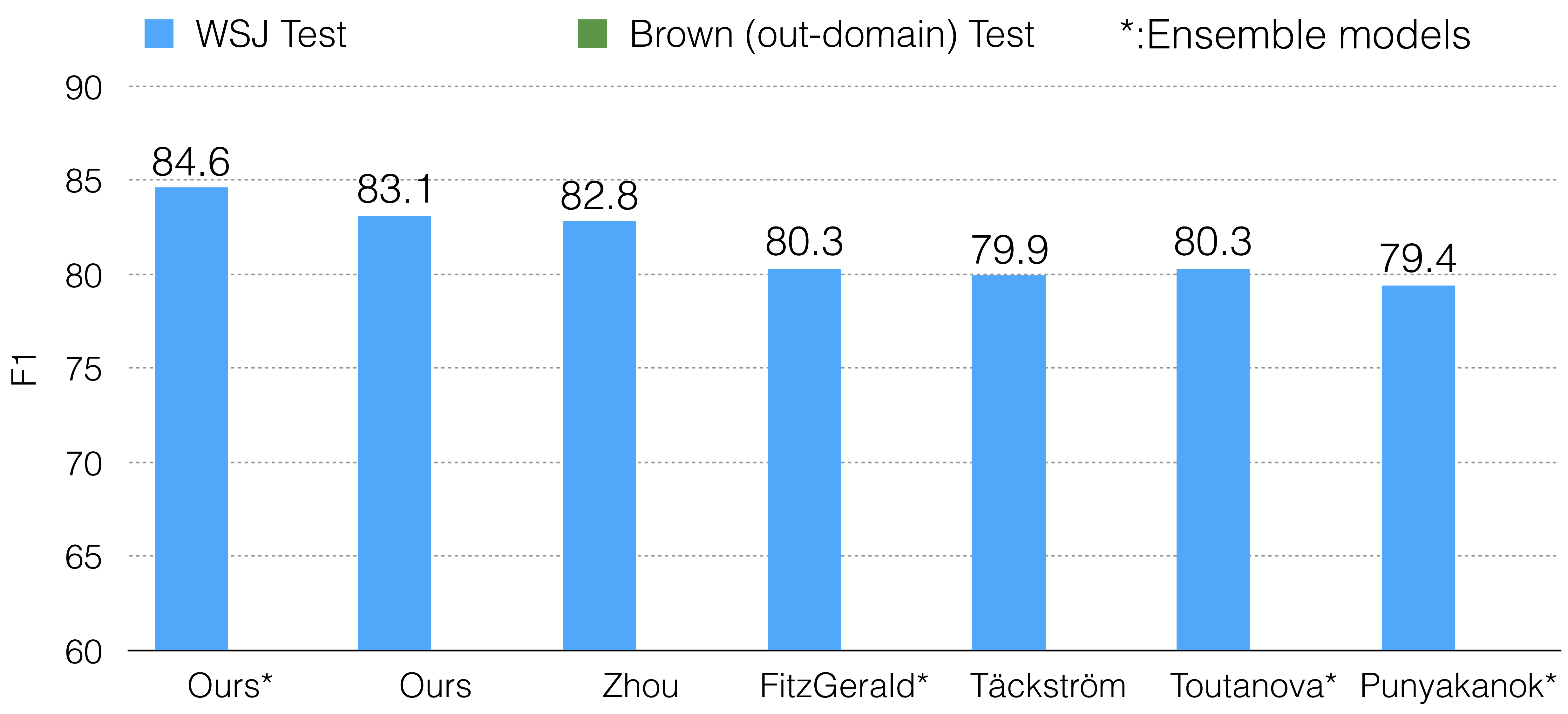
Datasets

	CoNLL-2005 (PropBank)	CoNLL-2012 (OntoNotes)
Size	40k sentences	140k sentences
Domains	newswire	• telephone conversations • newswire • newsgroups • broadcast news • broadcast conversation • weblogs
Annotated predicates	Verbs	Added some nominal predicates

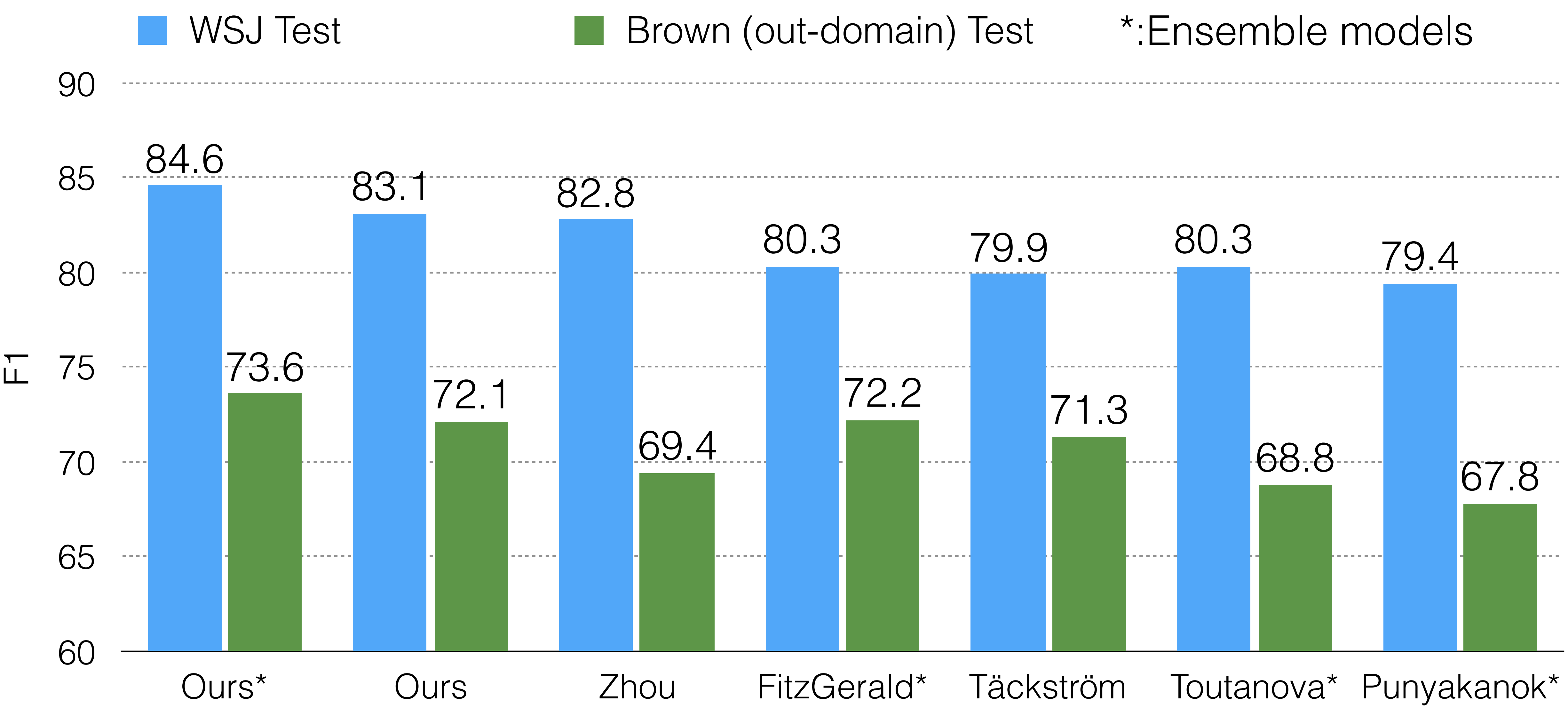
CoNLL 2005 Results



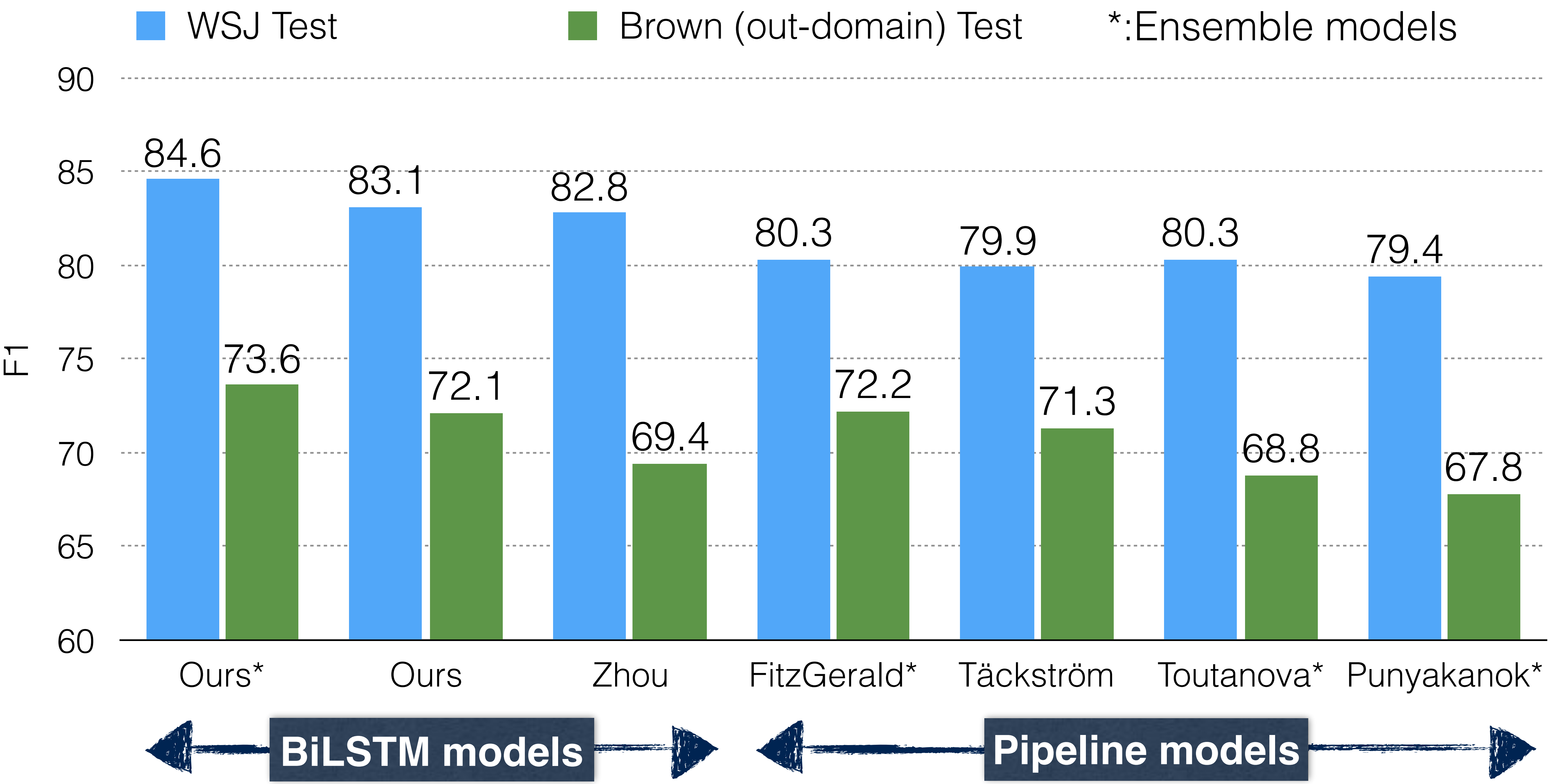
CoNLL 2005 Results



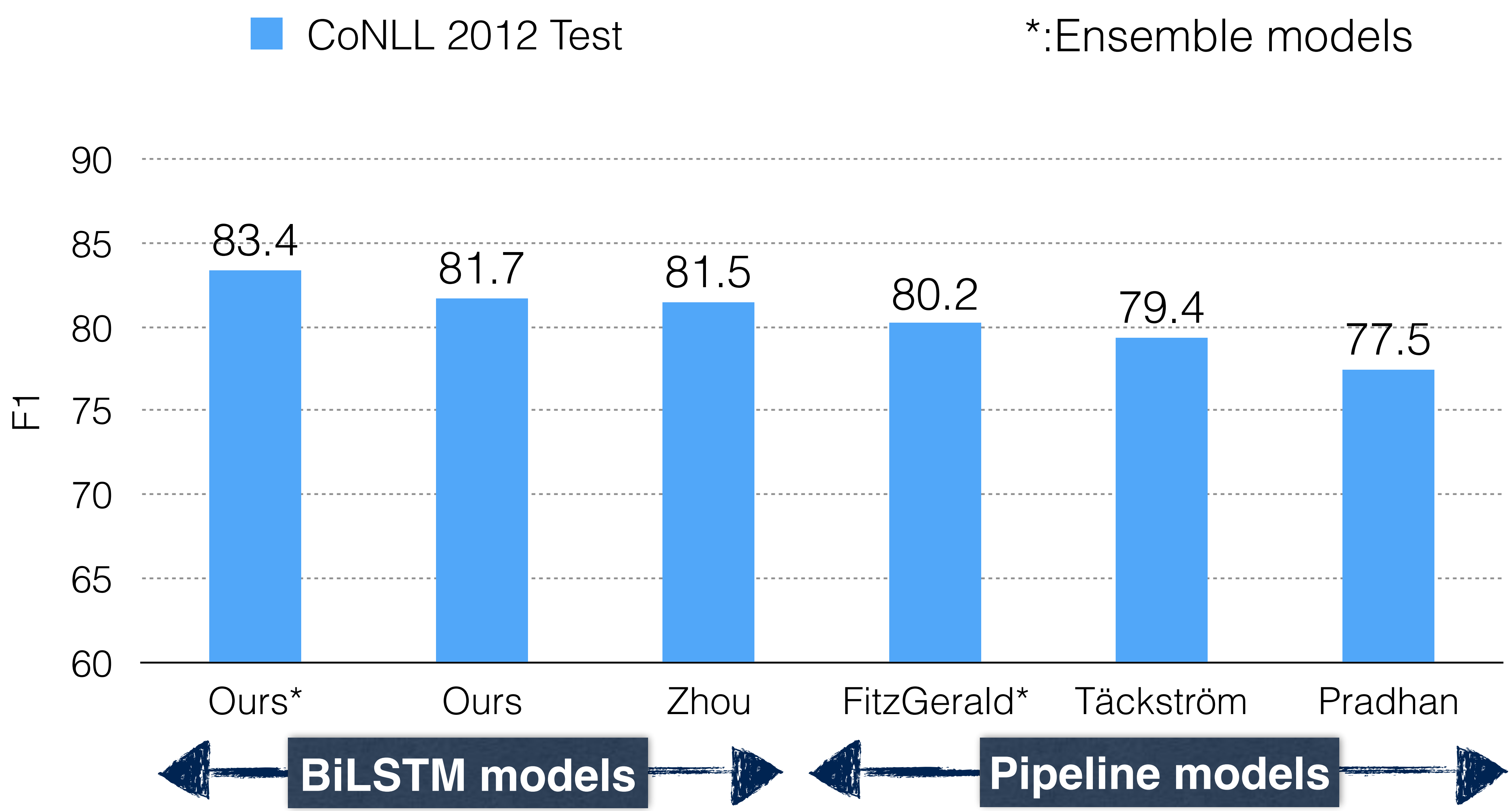
CoNLL 2005 Results



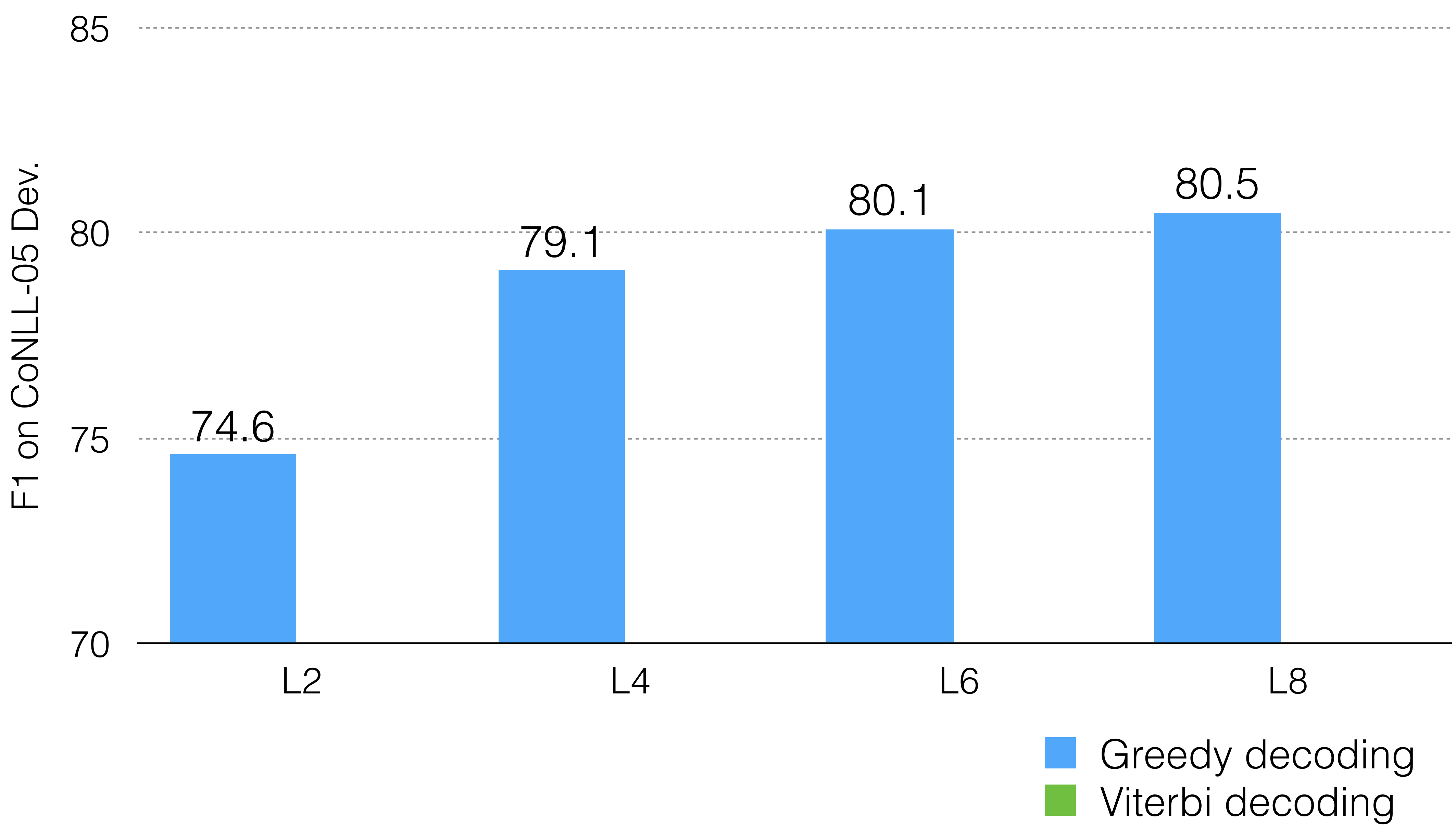
CoNLL 2005 Results



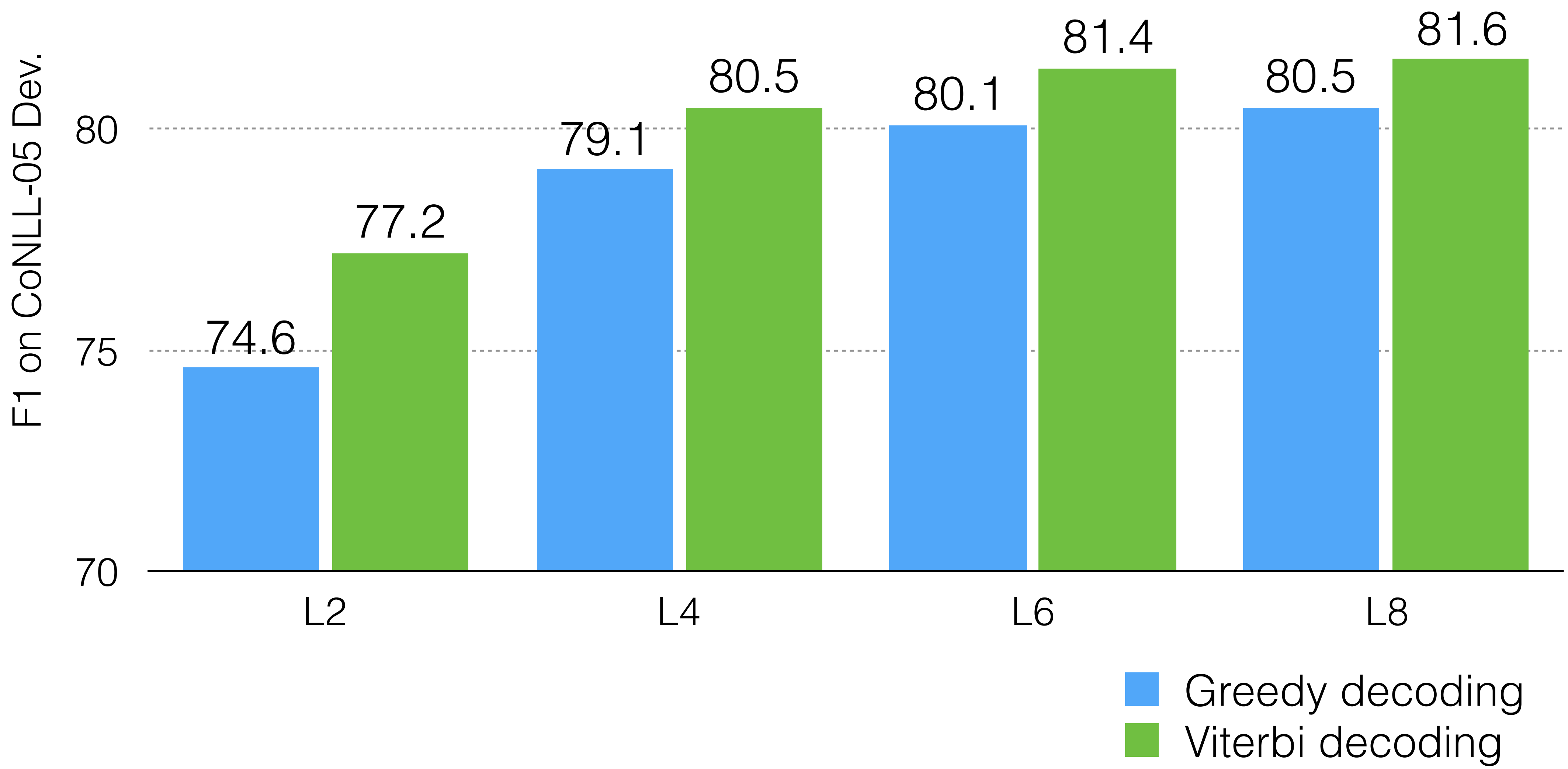
CoNLL 2012 (OntoNotes) Results



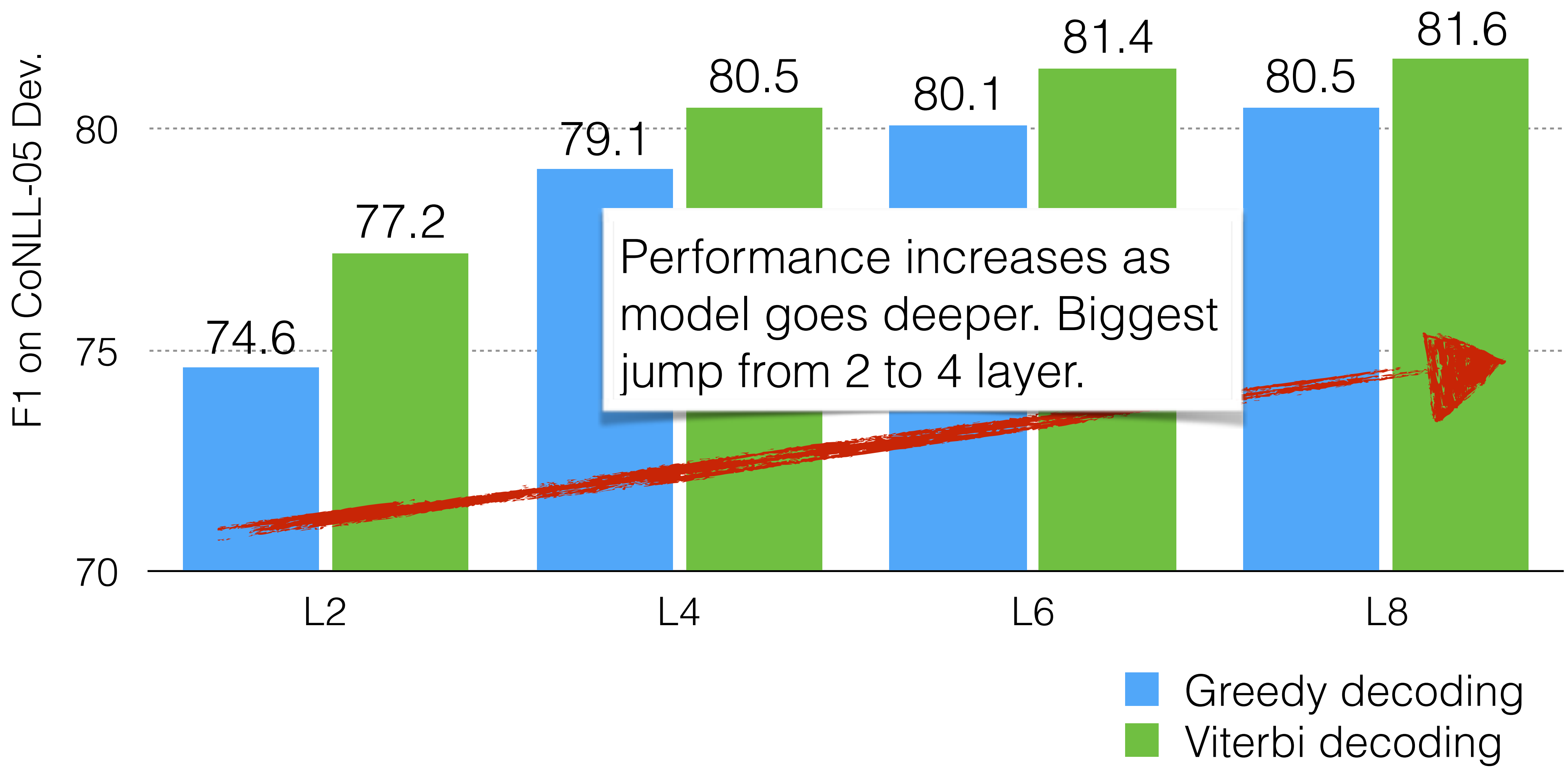
Ablations on Number of Layers (2,4,6 and 8)



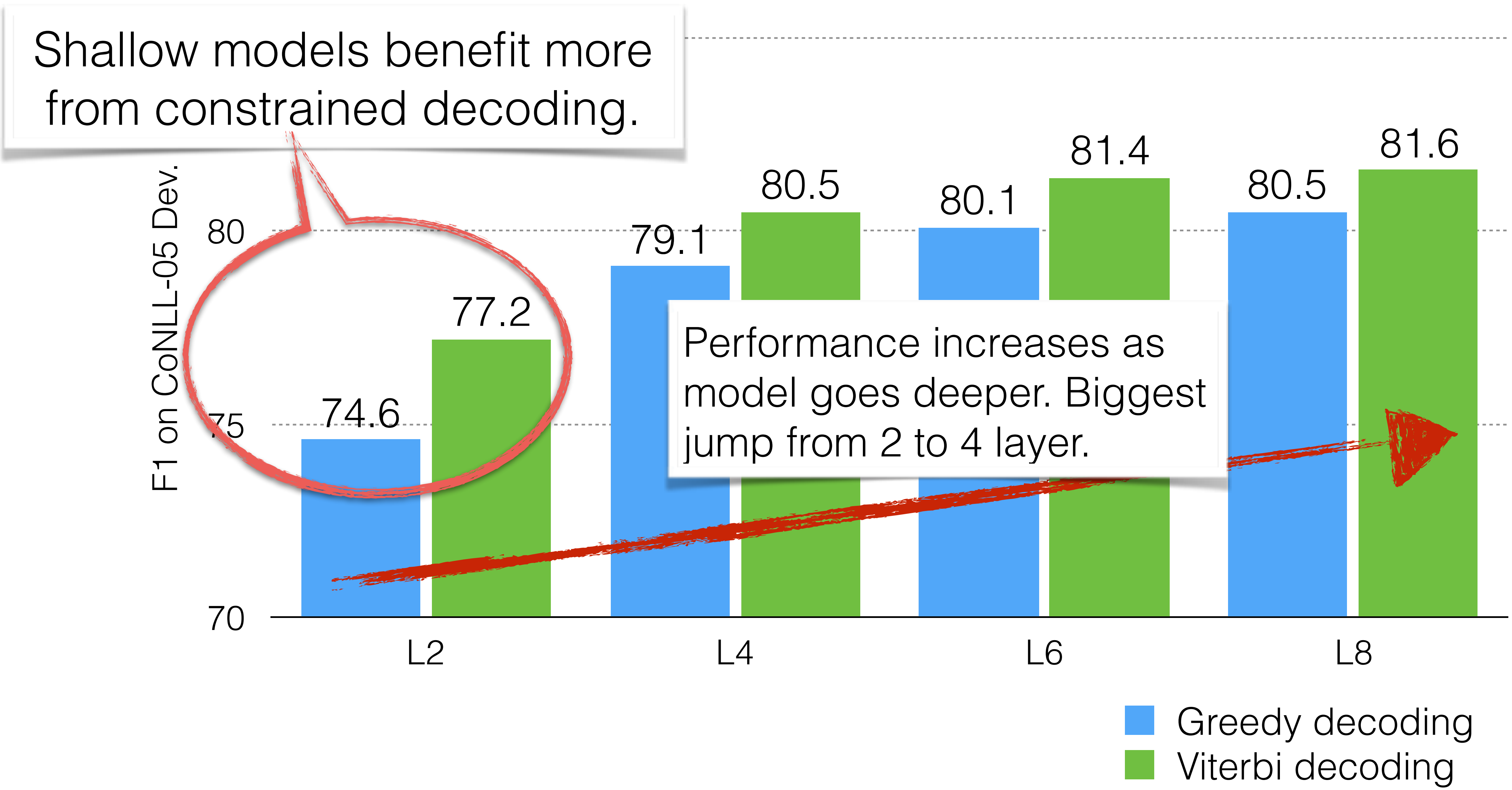
Ablations on Number of Layers (2,4,6 and 8)



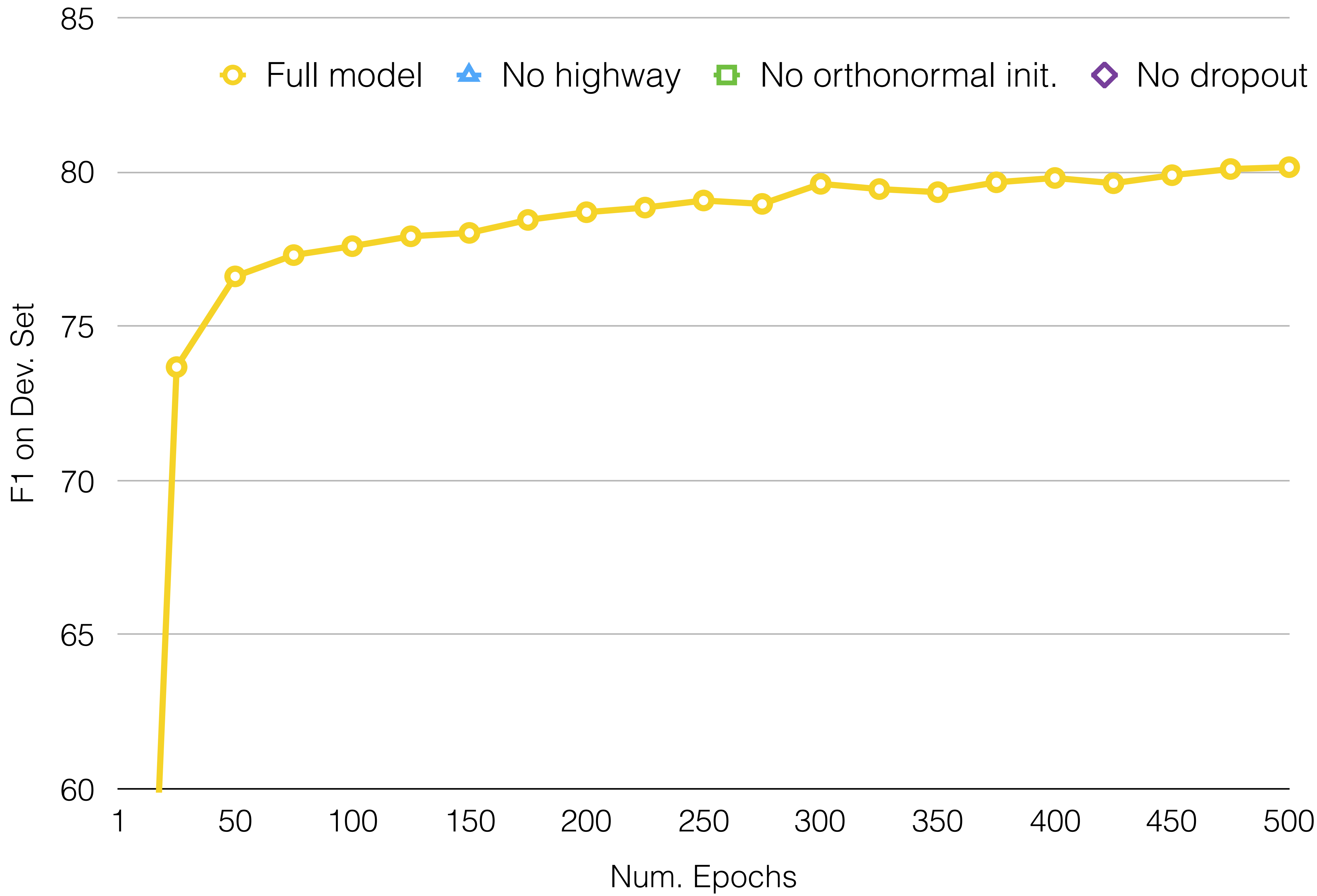
Ablations on Number of Layers (2,4,6 and 8)



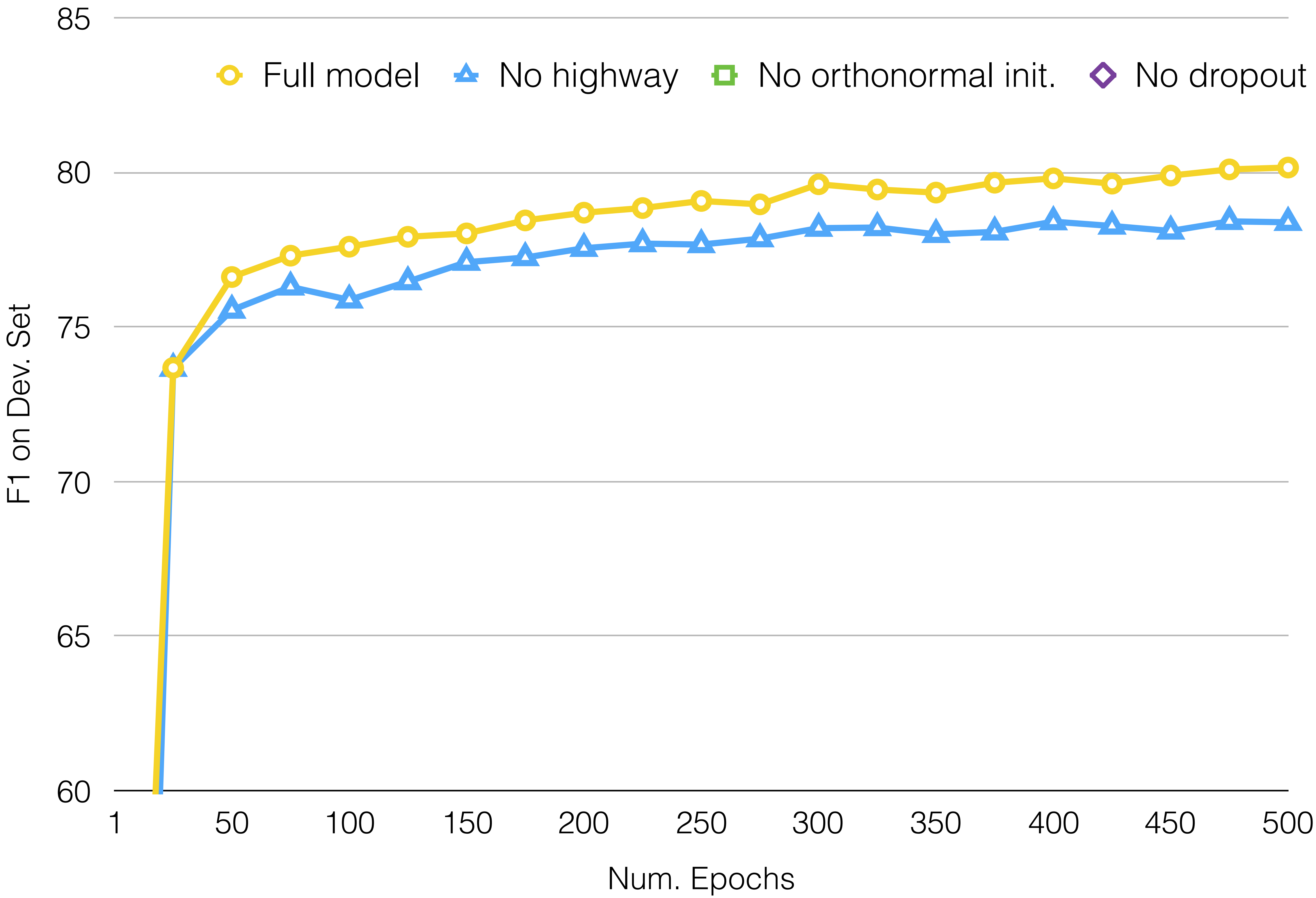
Ablations on Number of Layers (2,4,6 and 8)



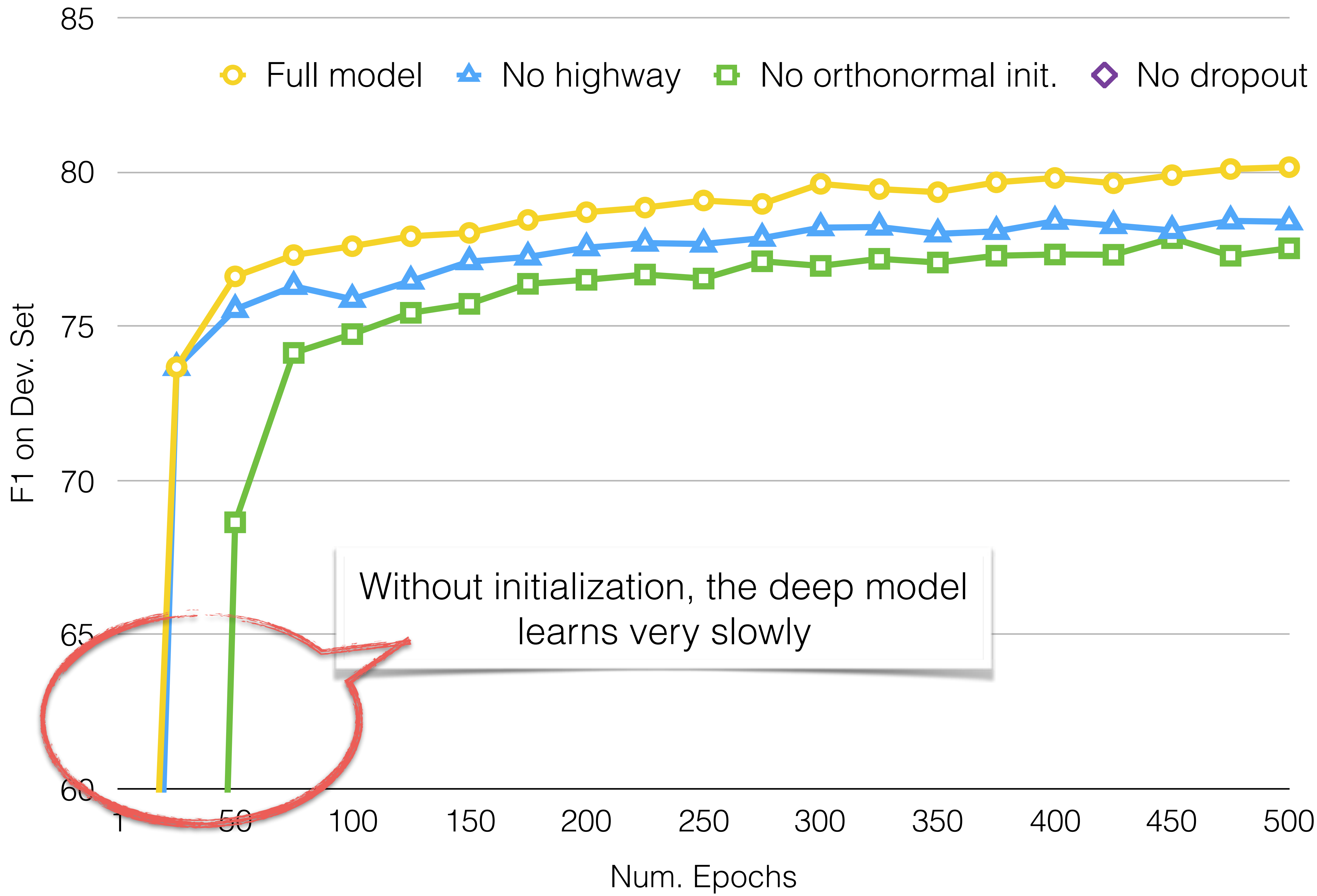
Ablations (single model, on CoNLL05 Dev)



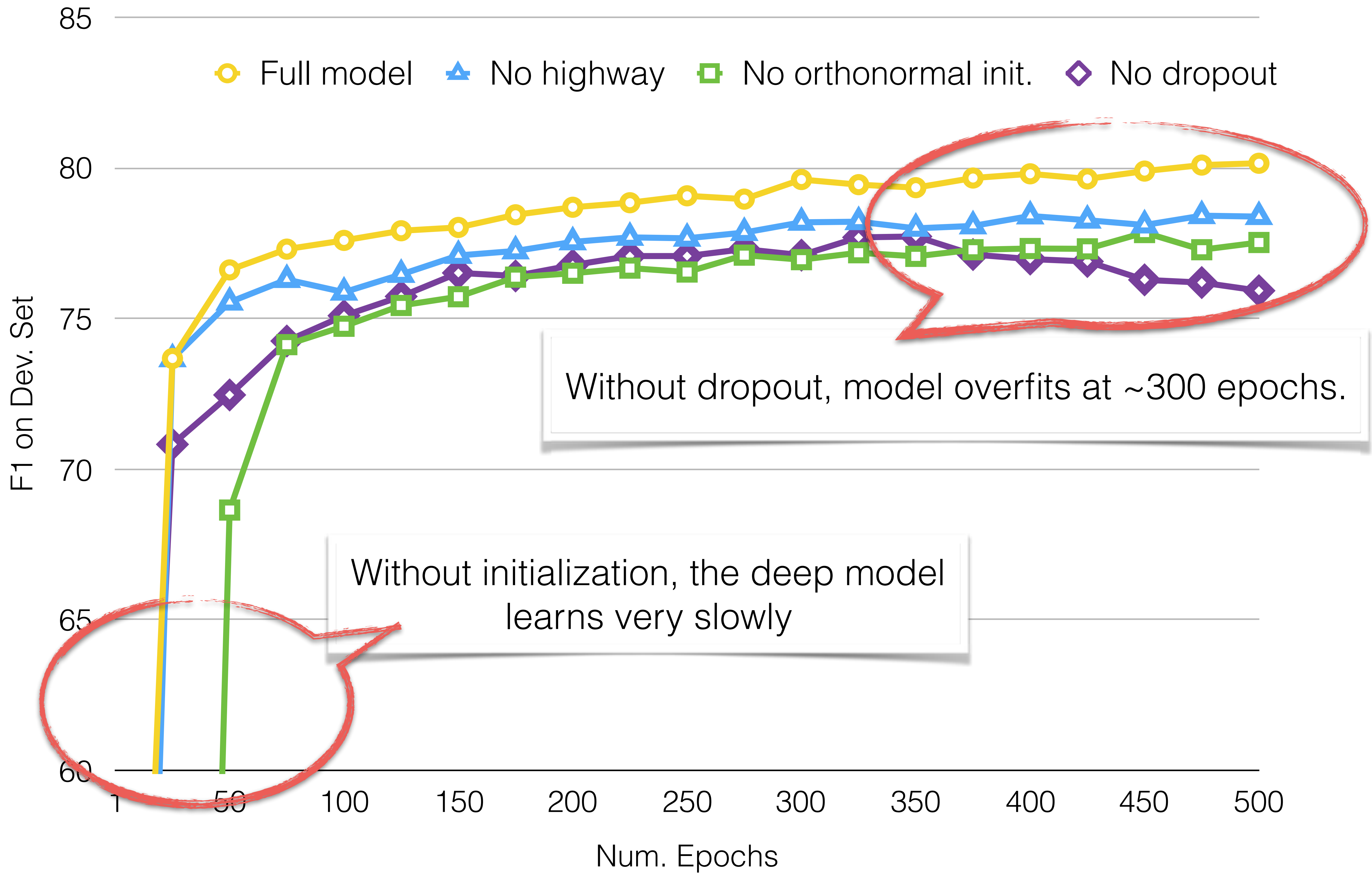
Ablations (single model, on CoNLL05 Dev)



Ablations (single model, on CoNLL05 Dev)

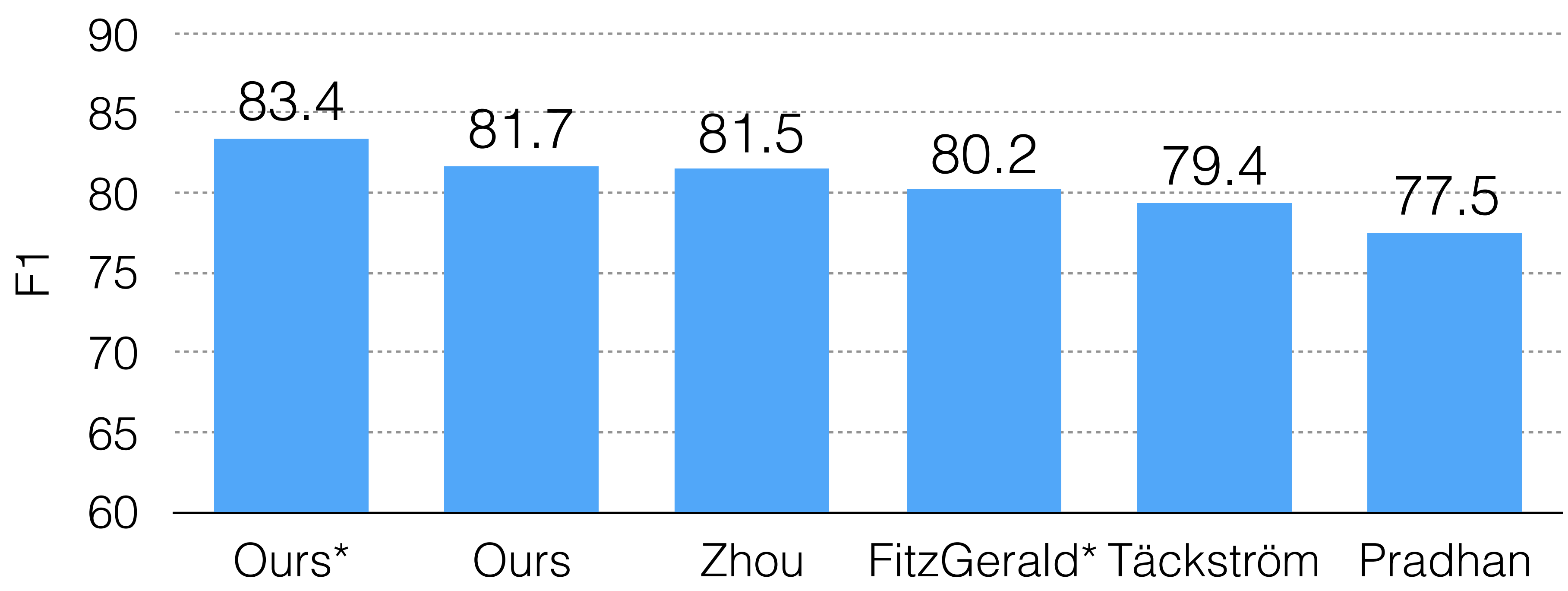


Ablations (single model, on CoNLL05 Dev)



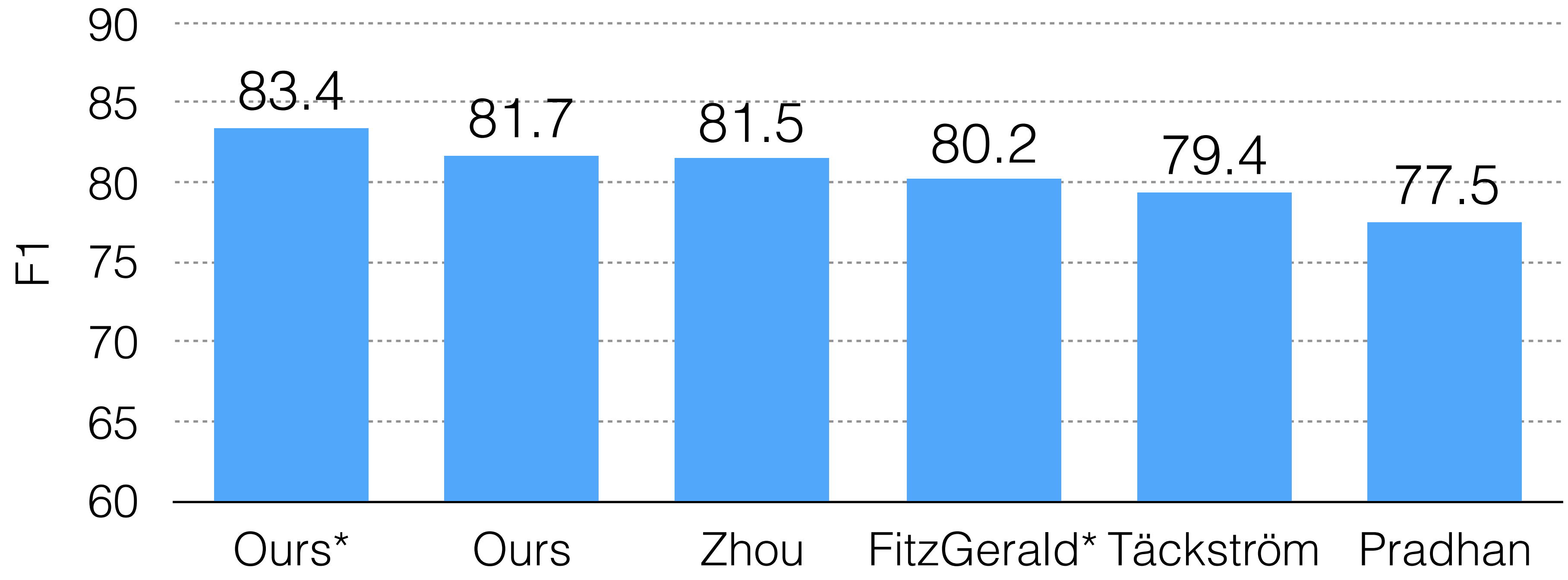
What can we learn from the results?

1. What's in the remaining 17%? When does the model still **struggle**?



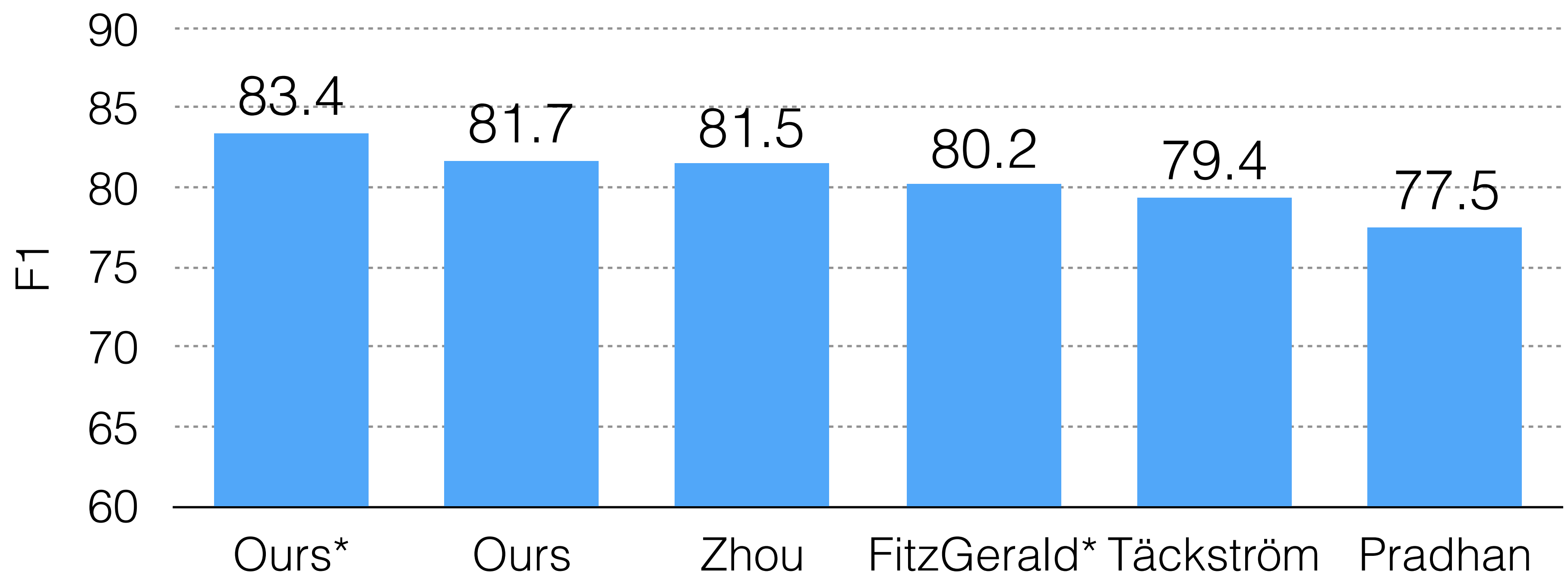
What can we learn from the results?

1. What's in the remaining 17%? When does the model still **struggle**?
2. What are **deeper models** good at?



What can we learn from the results?

1. What's in the remaining 17%? When does the model still **struggle**?
2. What are **deeper models** good at?
3. BiLSTM-based models are very accurate even without syntax. But can we conclude **syntax** is no longer useful in SRL?



Question (1): When does the model make mistakes?

Question (1): When does the model make mistakes?

Analysis

— Error breakdown with oracle transformation

— E.g. tease apart labeling errors and boundary errors

— Link the error types to known linguistic phenomena (e.g. pp attachment)

Error Breakdown

Oracle Transformations

Error Breakdown

Oracle Transformations

1) Fix Label:

[We] *fly* to NYC tomorrow.

~~ARG0~~

ARG1

Error Breakdown

Oracle Transformations

1) Fix Label:

[We] *fly* to NYC tomorrow.

~~ARG0~~

ARG1

2) Move core arg:

ARG0 ← ARG0

[They] wrote **[an email]** to *cancel* it.

Error Breakdown

Oracle Transformations

1) Fix Label:

[We] *fly* to NYC tomorrow.

~~ARG0~~

ARG1

2) Move core arg:

ARG0 ← ARG0

[They] wrote **[an email]** to *cancel* it.

3) Split/ Merge span:

ARG1

I *eat* **[pasta with delight]**.

ARG1 ARGM-MNR

[pasta] [with delight]

ARG1 ARGM-MNR

I *eat* **[pasta] [with broccoli]**.

ARG1

[pasta with broccoli]

Error Breakdown

Oracle Transformations

1) Fix Label:

[We] *fly* to NYC tomorrow.

~~ARG0~~

ARG1

2) Move core arg:

ARG0 ARG0

[They] wrote **[an email]** to *cancel* it.

3) Split/ Merge span:

ARG1

I *eat* **[pasta with delight]**.

ARG1 ARGM-MNR

[pasta] [with delight]

ARG1 ARGM-MNR

I *eat* **[pasta] [with broccoli]**.

ARG1

[pasta with broccoli]

4) Fix span boundary:

ARG1

[“No broccoli”,] I said.

[“No broccoli”] ,

Error Breakdown

Oracle Transformations

1) Fix Label:

[We] *fly* to NYC tomorrow.

~~ARG0~~

ARG1

2) Move core arg:

ARG0 ← ARG0

[They] wrote **[an email]** to *cancel* it.

3) Split/ Merge span:

ARG1

I *eat* **[pasta with delight]**.

ARG1 ARGM-MNR

[pasta] [with delight]

ARG1 ARGM-MNR

I *eat* **[pasta] [with broccoli]**.

ARG1

[pasta with broccoli]

4) Fix span boundary:

ARG1

[“No broccoli”,] I said.

[“No broccoli”] ,

5) Drop/ add arg:

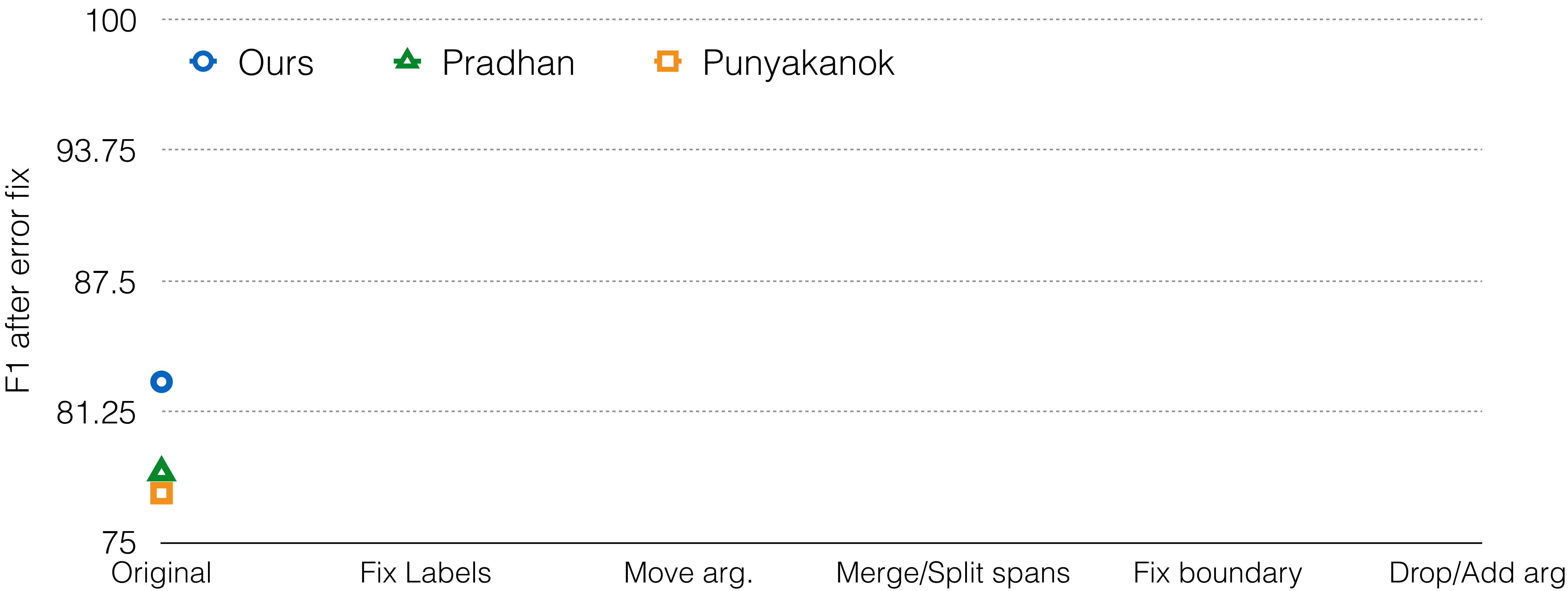
Hesitantly, they *declined* to elaborate **[on that matter]**.

~~ARG1~~

\+ ARGM-MNR

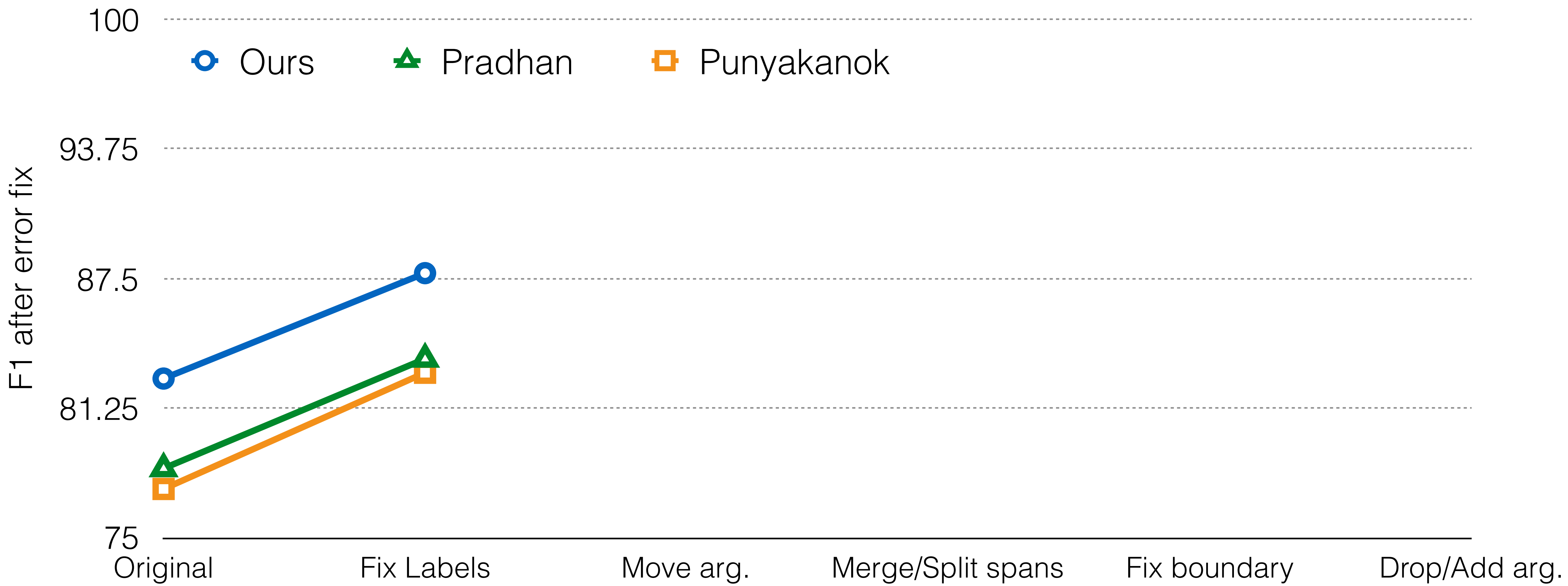
[Hesitantly], they *declined* to elaborate on that matter.

Error Breakdown



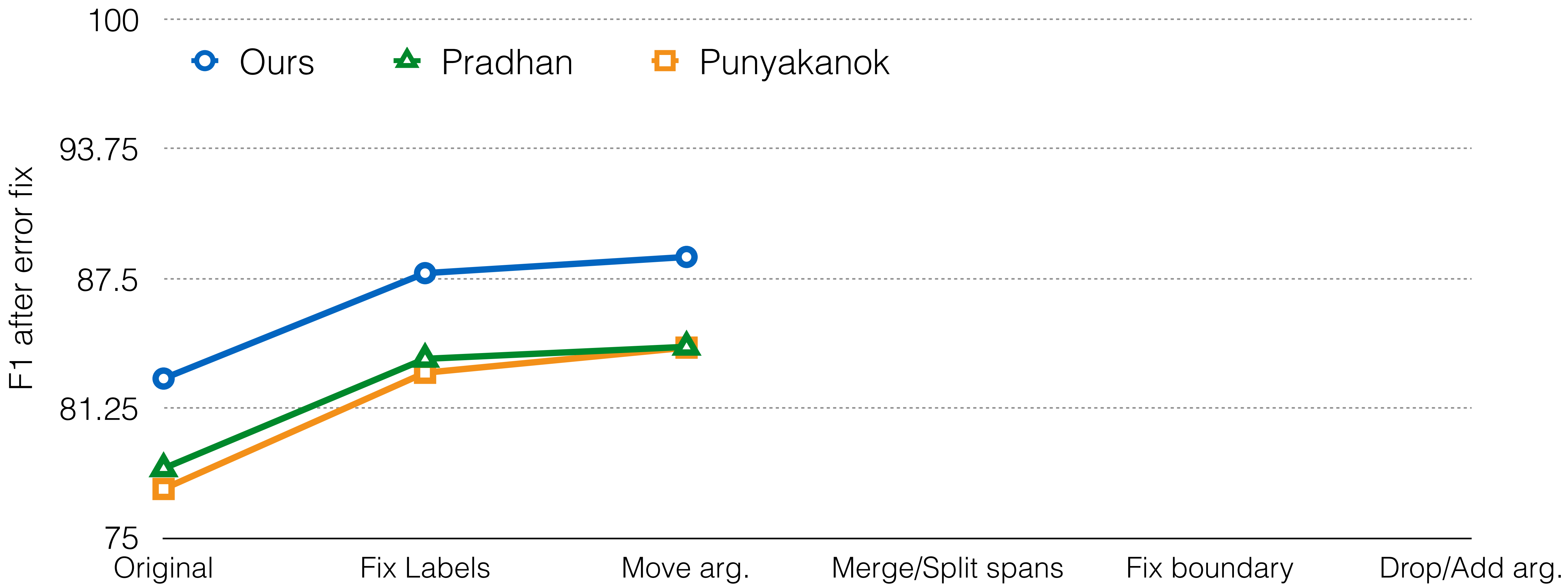
Pradhan, Punyakanok: CoNLL-2005 systems

Error Breakdown



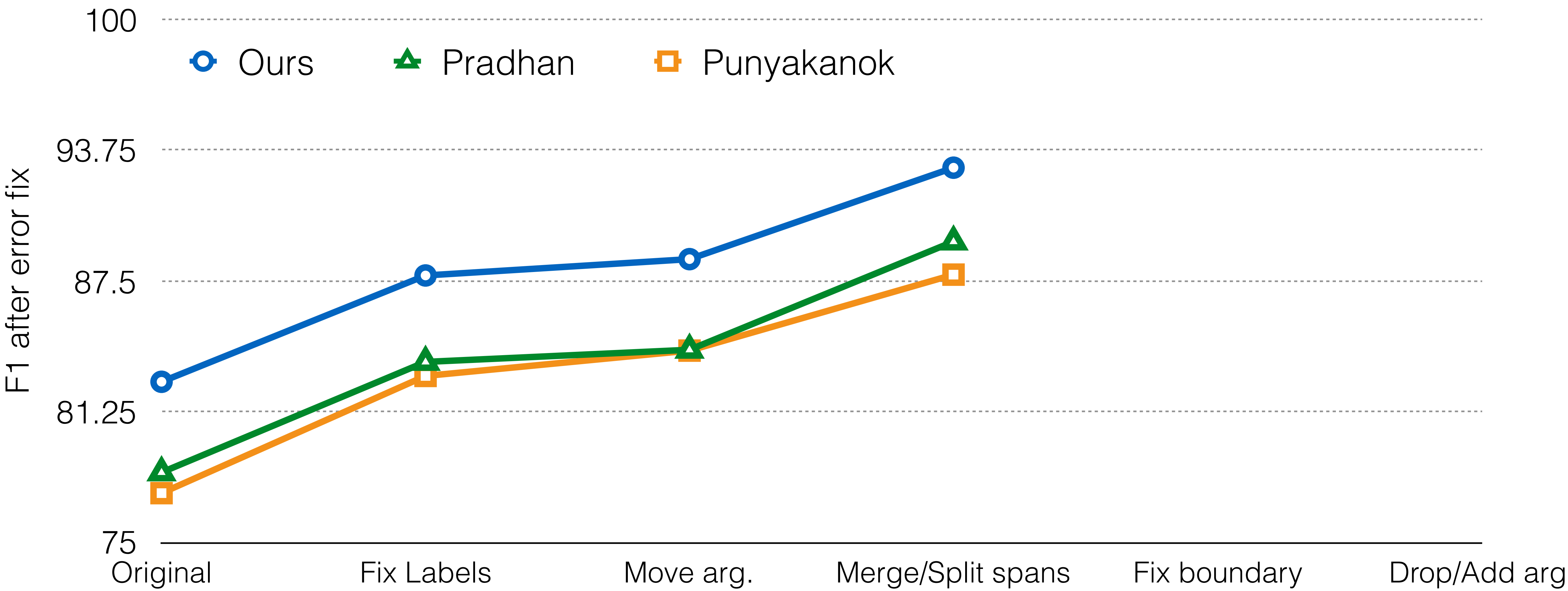
Pradhan, Punyakanok: CoNLL-2005 systems

Error Breakdown



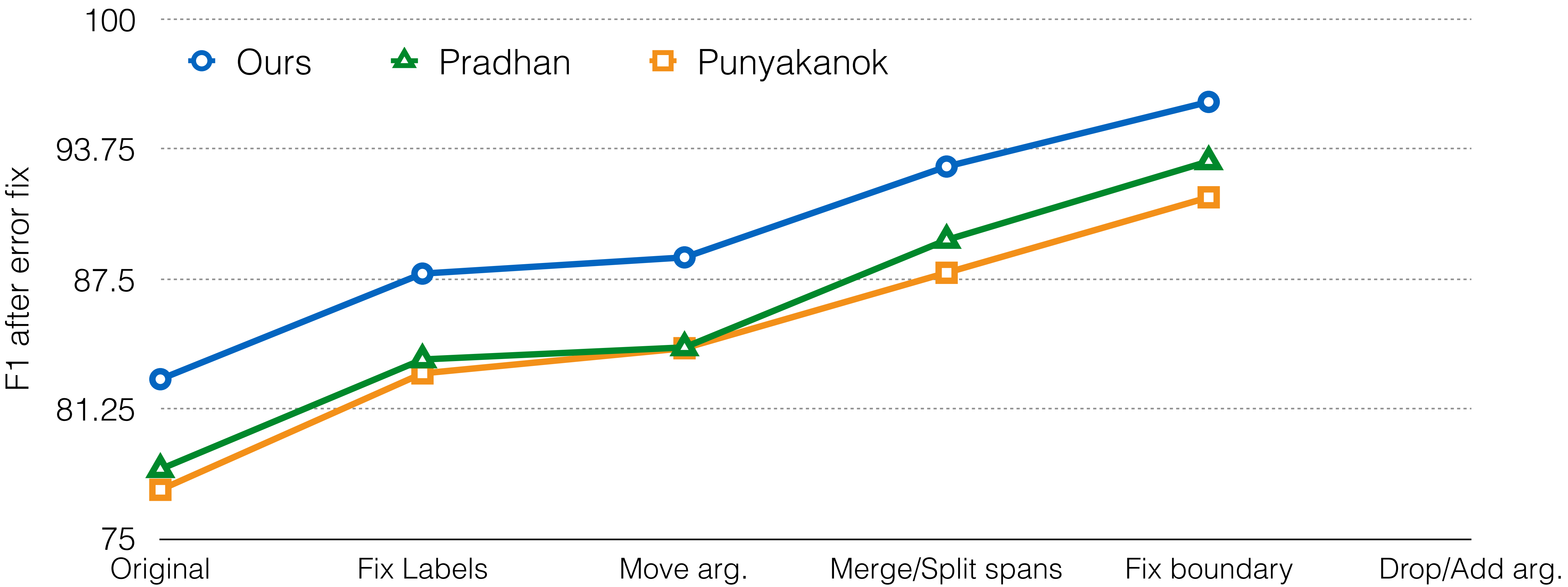
Pradhan, Punyakanok: CoNLL-2005 systems

Error Breakdown



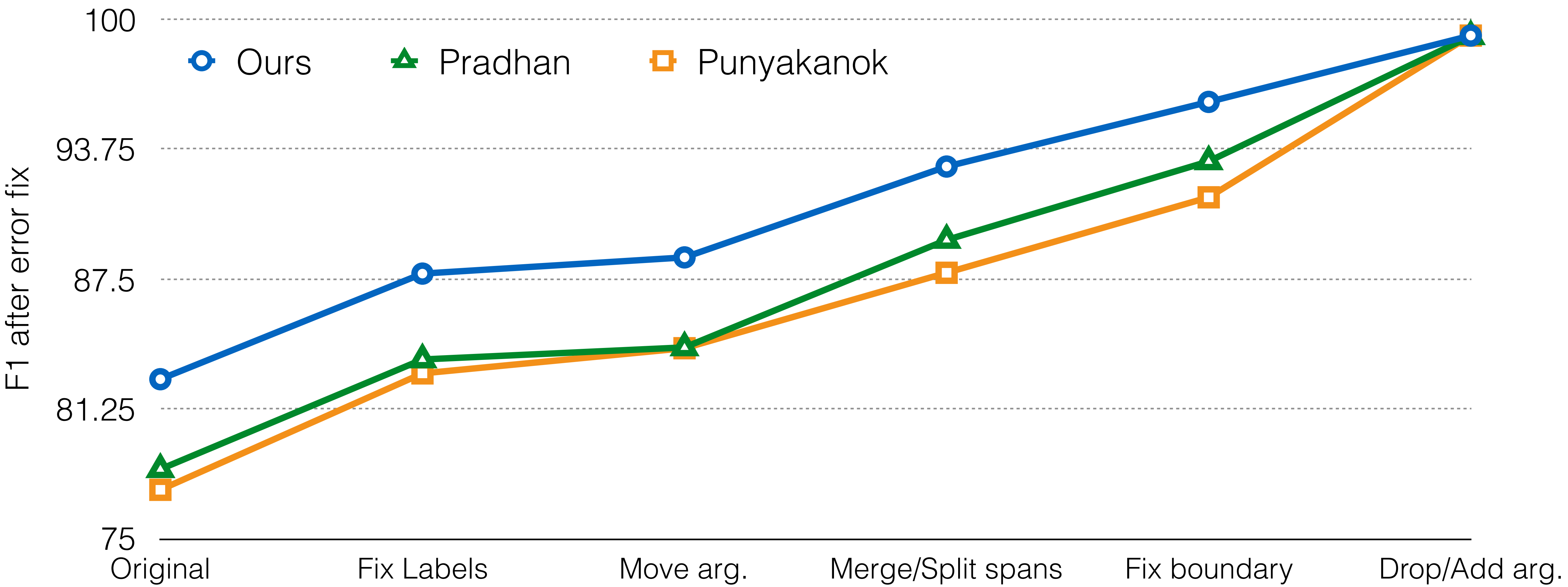
Pradhan, Punyakanok: CoNLL-2005 systems

Error Breakdown



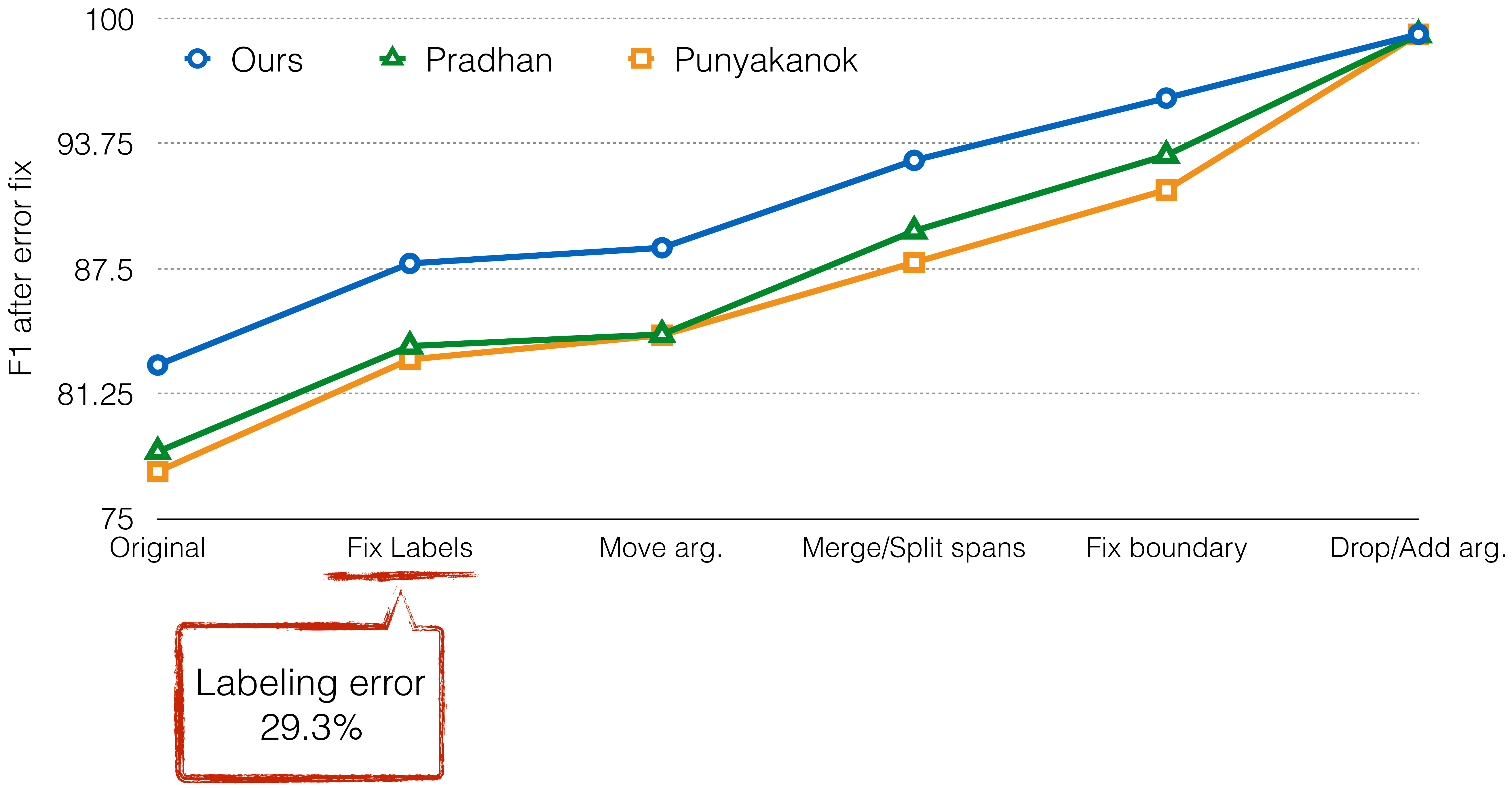
Pradhan, Punyakanok: CoNLL-2005 systems

Error Breakdown



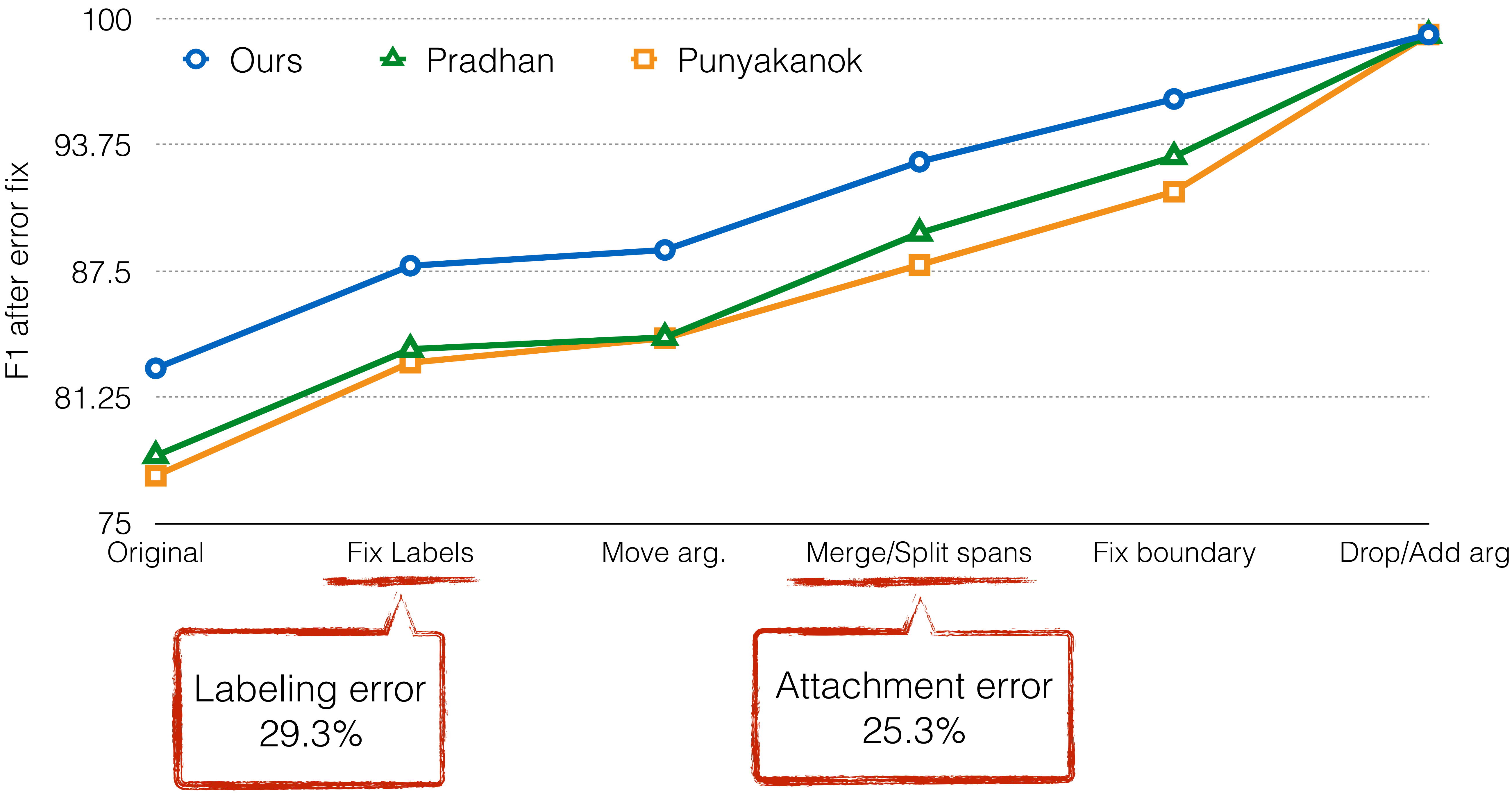
Pradhan, Punyakanok: CoNLL-2005 systems

Error Breakdown



Pradhan, Punyakanok: CoNLL-2005 systems

Error Breakdown



Pradhan, Punyakanok: CoNLL-2005 systems

Labeling Errors

Confusion matrix for labeling errors (row normalized)

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	76	13	6	14	2	0	0	0	0	0
A1	16	74	25	0	0	18	9	11	19	2
A2	2	5	31	52	10	45	26	46	19	0
A3	1	0	1	57	2	0	0	0	19	2
ADV	0	0	0	5	33	0	11	33	19	5
DIR	0	0	3	5	0	27	9	2	0	0
LOC	1	2	7	0	2	0	34	11	0	2
MNR	1	0	7	29	21	0	0	43	0	3
PNC	0	1	3	5	0	9	3	2	44	0
TMP	0	2	3	0	26	9	20	7	0	71

Labeling Errors

Confusion matrix for labeling errors (row normalized)

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	76	13	6	14	2	0	0	0	0	0
A1	16	74	25	0	0	18	9	11	19	2
A2	2	5	31	52	10	45	26	46	19	0
A3	1	0	1	57	2	0	0	0	19	2
ADV	0	0	0	5	33	0	11	33	19	5
DIR	0	0	3	5	0	27	9	2	0	0
LOC	1	2	7	0	2	0	34	11	0	2
MNR	1	0	7	29	21	0	0	43	0	3
PNC	0	1	3	5	0	9	3	2	44	0
TMP	0	2	3	0	26	9	20	7	0	71

- ARG2 is often confused with certain adjuncts (DIR, LOC, MNR), why?

Labeling Errors

Confusion matrix for labeling errors (row normalized)

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	76	13	6	14	2	0	0	0	0	0
A1	16	74	25	0	0	18	9	11	19	2
A2	2	5	31	52	10	45	26	46	19	0
A3	1	0	1	57	2	0	0	0	19	2
ADV	0	0	0	5	33	0	11	33	19	5
DIR	0	0	3	5	0	27	9	2	0	0
LOC	1	2	7	0	2	0	34	11	0	2
MNR	1	0	7	29	21	0	0	43	0	3
PNC	0	1	3	5	0	9	3	2	44	0
TMP	0	2	3	0	26	9	20	7	0	71

- ARG2 is often confused with certain adjuncts (DIR, LOC, MNR), why?

Predicate: ***move***

Arg0-PAG: *mover*
Arg1-PPT: *moved*
Arg2-GOL: *destination*
Arg3-VSP: *aspect, domain in which arg1 moving*

Labeling Errors

Confusion matrix for labeling errors (row normalized)

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	76	13	6	14	2	0	0	0	0	0
A1	16	74	25	0	0	18	9	11	19	2
A2	2	5	31	52	10	45	26	46	19	0
A3	1	0	1	57	2	0	0	0	19	2
ADV	0	0	0	5	33	0	11	33	19	5
DIR	0	0	3	5	0	27	9	2	0	0
LOC	1	2	7	0	2	0	34	11	0	2
MNR	1	0	7	29	21	0	0	43	0	3
PNC	0	1	3	5	0	9	3	2	44	0
TMP	0	2	3	0	26	9	20	7	0	71

- ARG2 is often confused with certain adjuncts (DIR, LOC, MNR), why?

Predicate: ***move***

Arg0-PAG: *mover*
Arg1-PPT: *moved*
Arg2-GOL: *destination*
Arg3-VSP: *aspect, domain in which arg1 moving*

Predicate: ***cut***

Arg0-PAG: *intentional cutter*
Arg1-PPT: *thing cut*
Arg2-DIR: *medium, source*
Arg3-MNR: *instrument, unintentional cutter*
Arg4-GOL: *beneficiary*

Labeling Errors

Confusion matrix for labeling errors (row normalized)

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	76	13	6	14	2	0	0	0	0	0
A1	16	74	25	0	0	18	9	11	19	2
A2	2	5	31	52	10	45	26	46	19	0
A3	1	0	1	57	2	0	0	0	19	2
ADV	0	0	0	5	33	0	11	33	19	5
DIR	0	0	3	5	0	27	9	2	0	0
LOC	1	2	7	0	2	0	34	11	0	2
MNR	1	0	7	29	21	0	0	43	0	3
PNC	0	1	3	5	0	9	3	2	44	0
TMP	0	2	3	0	26	9	20	7	0	71

- ARG2 is often confused with certain adjuncts (DIR, LOC, MNR), why?

Predicate: ***move***

Arg0-PAG: *mover*
Arg1-PPT: *moved*
Arg2-GOL: *destination*
Arg3-VSP: *aspect, domain in which arg1 moving*

Predicate: ***cut***

Arg0-PAG: *intentional cutter*
Arg1-PPT: *thing cut*
Arg2-DIR: *medium, source*
Arg3-MNR: *instrument, unintentional cutter*
Arg4-GOL: *beneficiary*

Predicate: ***strike***

Arg0-PAG: *Agent*
Arg1-PPT: *Theme(-Creation)*
Arg2-MNR: *Instrument*

Labeling Errors

Confusion matrix for labeling errors (row normalized)

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	76	13	6	14	2	0	0	0	0	0
A1	16	74	25	0	0	18	9	11	19	2
A2	2	5	31	52	10	45	26	46	19	0
A3	1	0	1	57	2	0	0	0	19	2
ADV	0	0	0	5	33	0	11	33	19	5
DIR	0	0	3	5	0	27	9	2	0	0
LOC	1	2	7	0	2	0	34	11	0	2
MNR	1	0	7	29	21	0	0	43	0	3
PNC	0	1	3	5	0	9	3	2	44	0
TMP	0	2	3	0	26	9	20	7	0	71

- ARG2 is often confused with certain adjuncts (DIR, LOC, MNR), why?

Predicate: ***move***

Arg0-PAG: *mover*
Arg1-PPT: *moved*
Arg2-GOL: *destination*
Arg3-VSP: *aspect, domain in which arg1 moving*

Predicate: ***cut***

Arg0-PAG: *intentional cutter*
Arg1-PPT: *thing cut*
Arg2-DIR: *medium, source*
Arg3-MNR: *instrument, unintentional cutter*
Arg4-GOL: *beneficiary*

Predicate: ***strike***

Arg0-PAG: *Agent*
Arg1-PPT: *Theme(-Creation)*
Arg2-MNR: *Instrument*

- Argument-adjunct distinctions are difficult even for human annotators!

Labeling Errors

Confusion matrix for labeling errors (row normalized)

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	76	13	6	14	2	0	0	0	0	0
A1	16	74	25	0	0	18	9	11	19	2
A2	2	5	31	52	10	45	26	46	19	0
A3	1	0	1	57	2	0	0	0	19	2
ADV	0	0	0	5	33	0	11	33	19	5
DIR	0	0	3	5	0	27	9	2	0	0
LOC	1	2	7	0	2	0	34	11	0	2
MNR	1	0	7	29	21	0	0	43	0	3
PNC	0	1	3	5	0	9	3	2	44	0
TMP	0	2	3	0	26	9	20	7	0	71

- ARG2 is often confused with certain adjuncts (DIR, LOC, MNR), why?

"After many attempts to find a reliable test to distinguish between arguments and adjuncts, we abandoned structurally marking this difference."

—The Penn Treebank: An Overview (Taylor et al., 2003)

- Argument-adjunct distinctions are difficult even for human annotators!

PP Attachment

Sumimoto ***financed*** the acquisition from Sears

PP Attachment

PP Attachment

Wrong PP attachment (attach high)

Arg1 (NP) Arg2 (PP)

Sumimoto ***financed*** the acquisition from Sears

Arg1 (NP)

Correct PP attachment (attach low)

Wrong SRL spans

merge

Correct SRL spans

PP Attachment

Wrong PP attachment
(attach high)

Arg1 (NP) Arg2 (PP)

Sumimoto ***financed*** the acquisition from Sears

Arg1 (NP)

Correct PP attachment
(attach low)

Wrong SRL spans

merge

Correct SRL spans

Merge/split span operations: 25.3%.
of the mode mistakes.

Categorize the Y spans in :
[XY]—>[X]**[Y]** and
[X]**[Y]**—>[XY] operations
using gold syntactic labels

PP Attachment

Wrong PP attachment
(attach high)

Arg1 (NP) Arg2 (PP)

Sumimoto ***financed*** the acquisition from Sears

Arg1 (NP)

Correct PP attachment
(attach low)

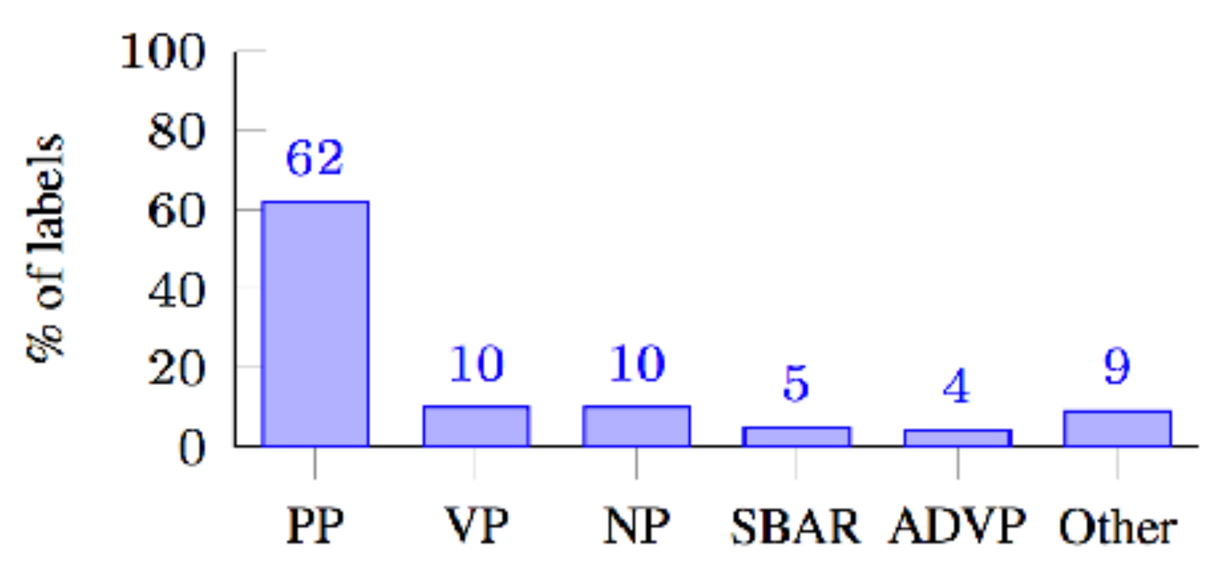
Wrong SRL spans

merge

Correct SRL spans

Merge/split span operations: 25.3%.
of the mode mistakes.

Categorize the Y spans in :
[XY]—>[X]**[Y]** and
[X]**[Y]**—>[XY] operations
using gold syntactic labels



Question (1): When does the model make mistakes?

Analysis

— Error breakdown with oracle transformation

— E.g. tease apart labeling errors and boundary errors

— Link the error types to known linguistic phenomena
(e.g. pp attachment)

Question (1): When does the model make mistakes?

Analysis

— Error breakdown with oracle transformation

— E.g. tease apart labeling errors and boundary errors

— Link the error types to known linguistic phenomena (e.g. pp attachment)

Takeaway

— Traditionally hard tasks, such as argument-adjunct distinction and PP attachment decisions are still challenging!

— Use external information to improve PP attachment.

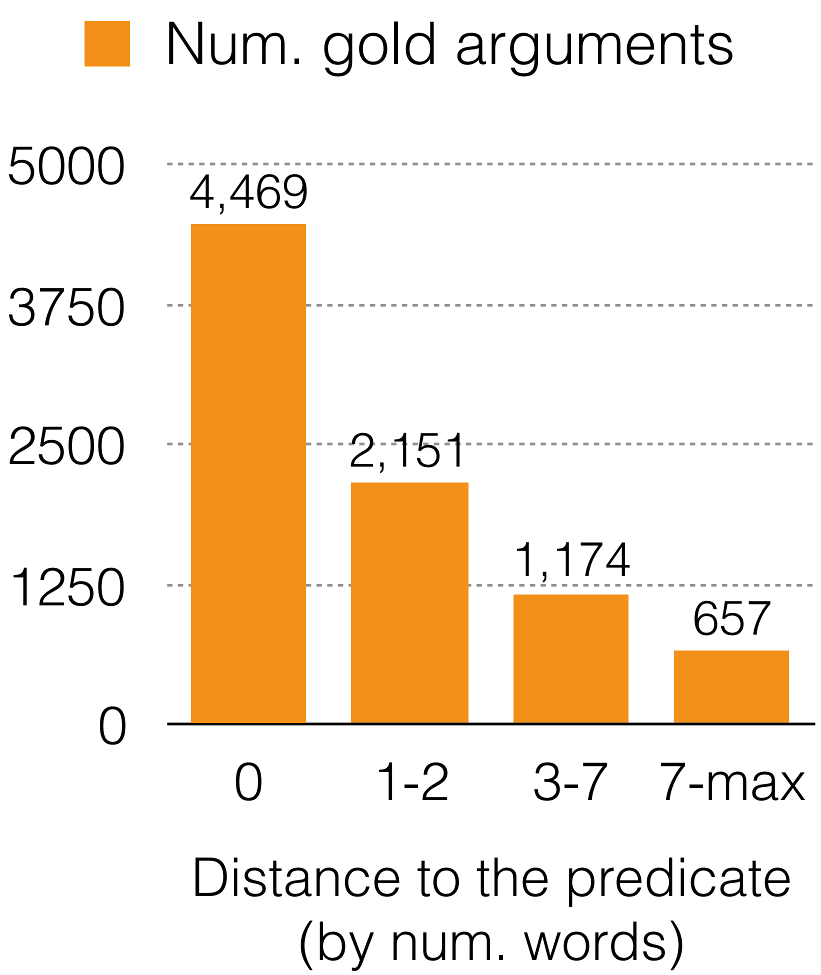
Question (2): What are deeper models good at?

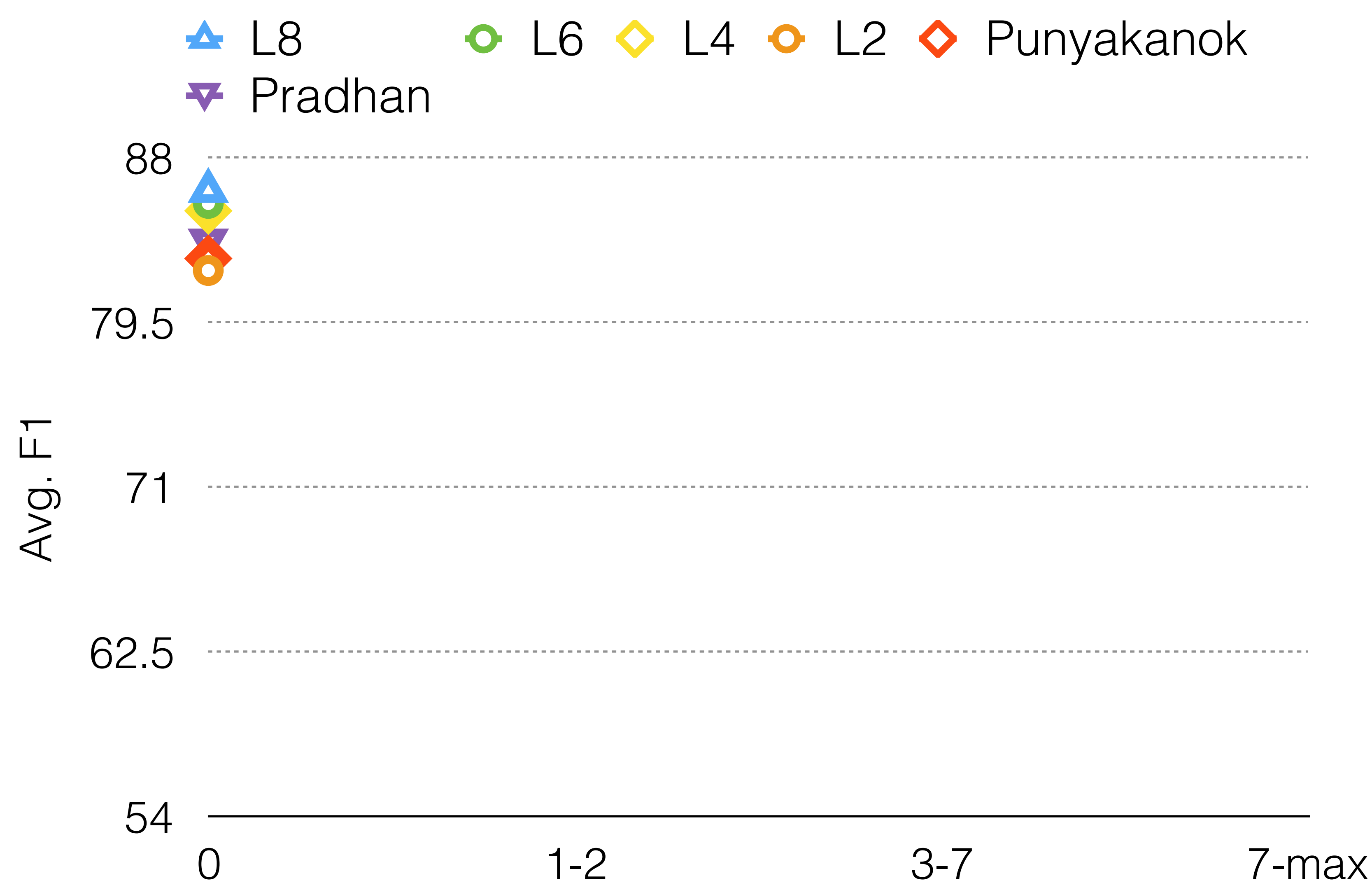
Analysis

— Long-range dependencies: model performance on arguments that are far away from the predicates.

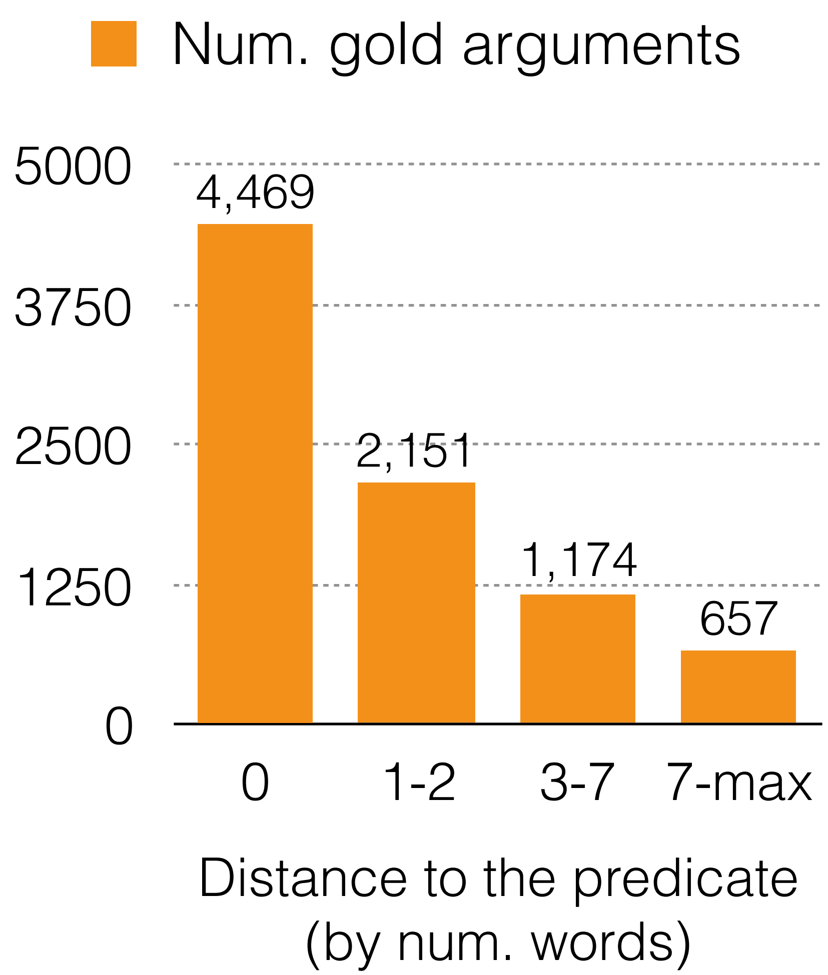
— Structural consistency: amount of inconsistent BIO tag pairs in greedy prediction.

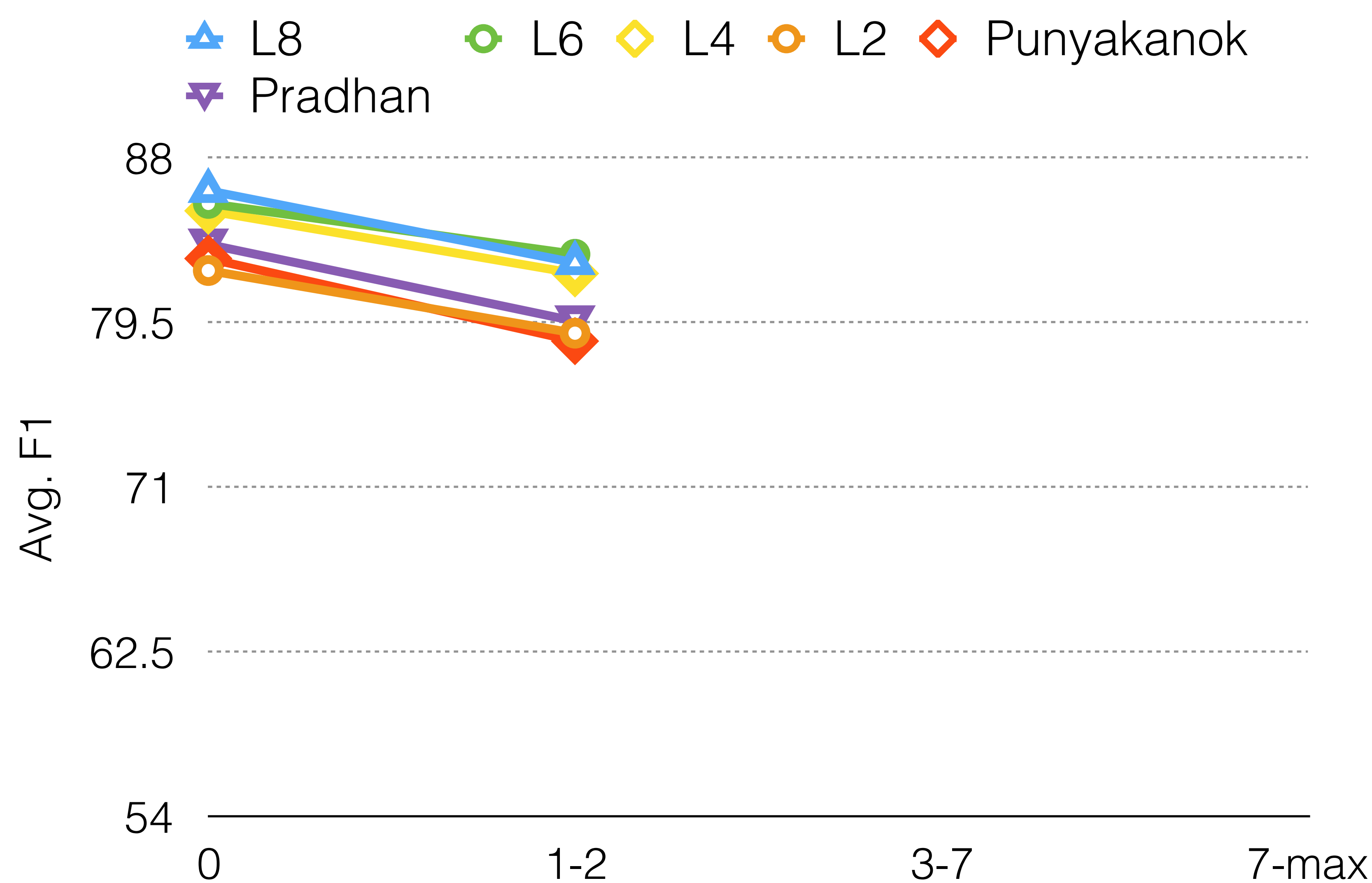
Long-range Dependencies



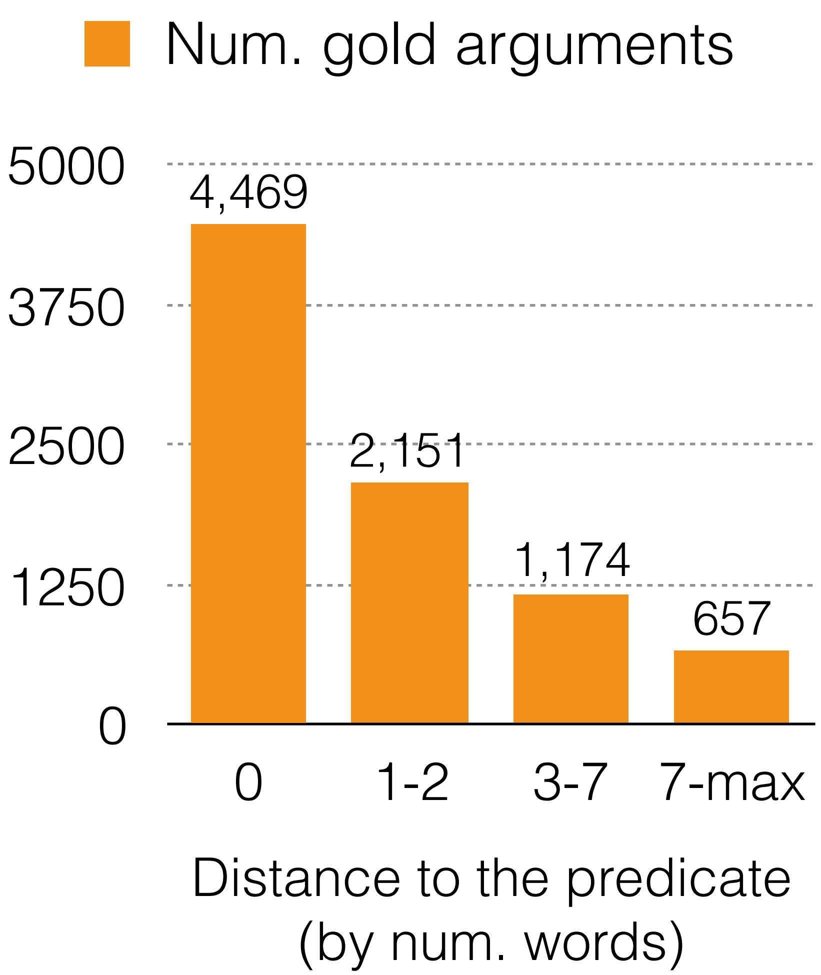


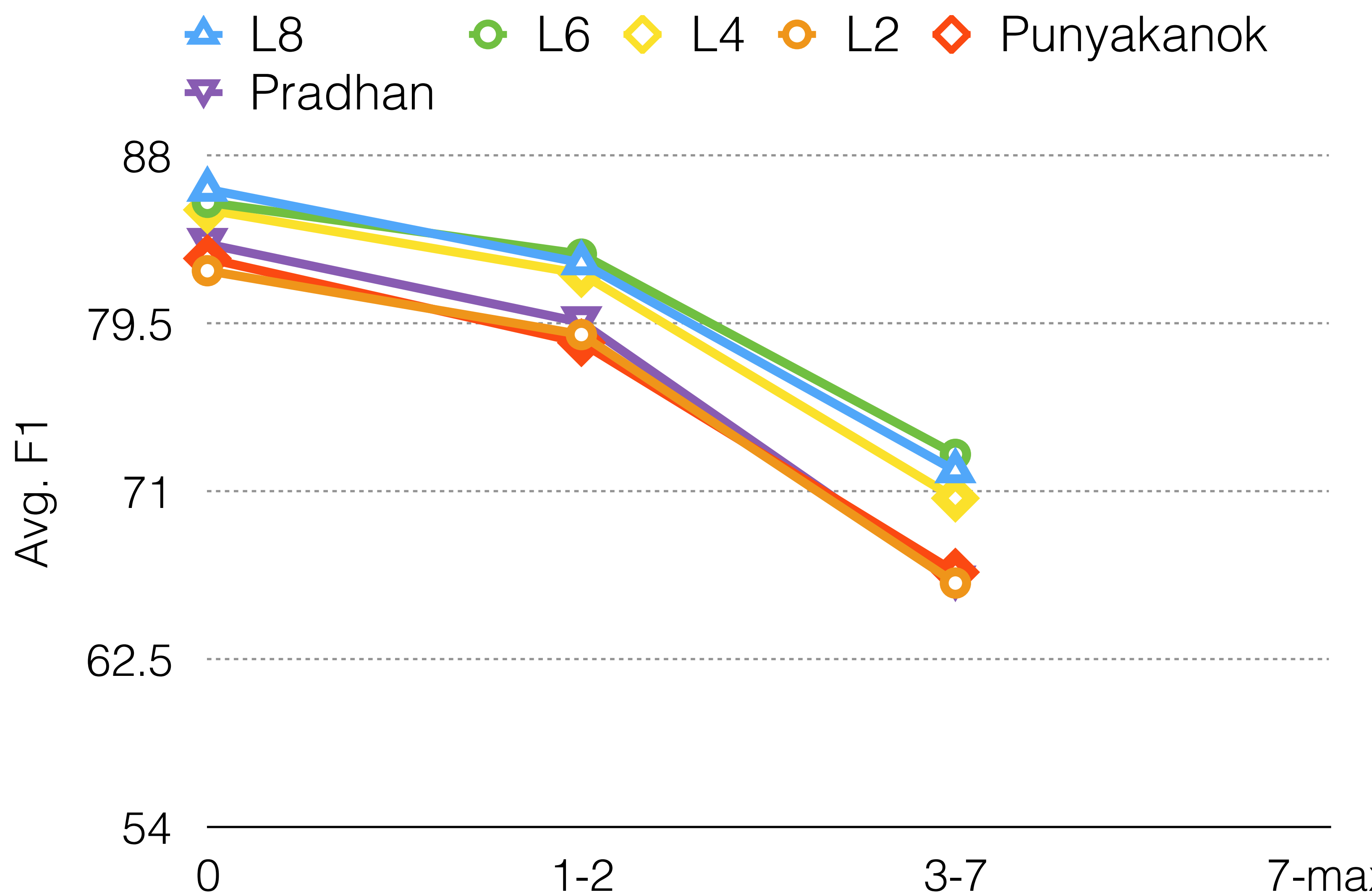
Long-range Dependencies



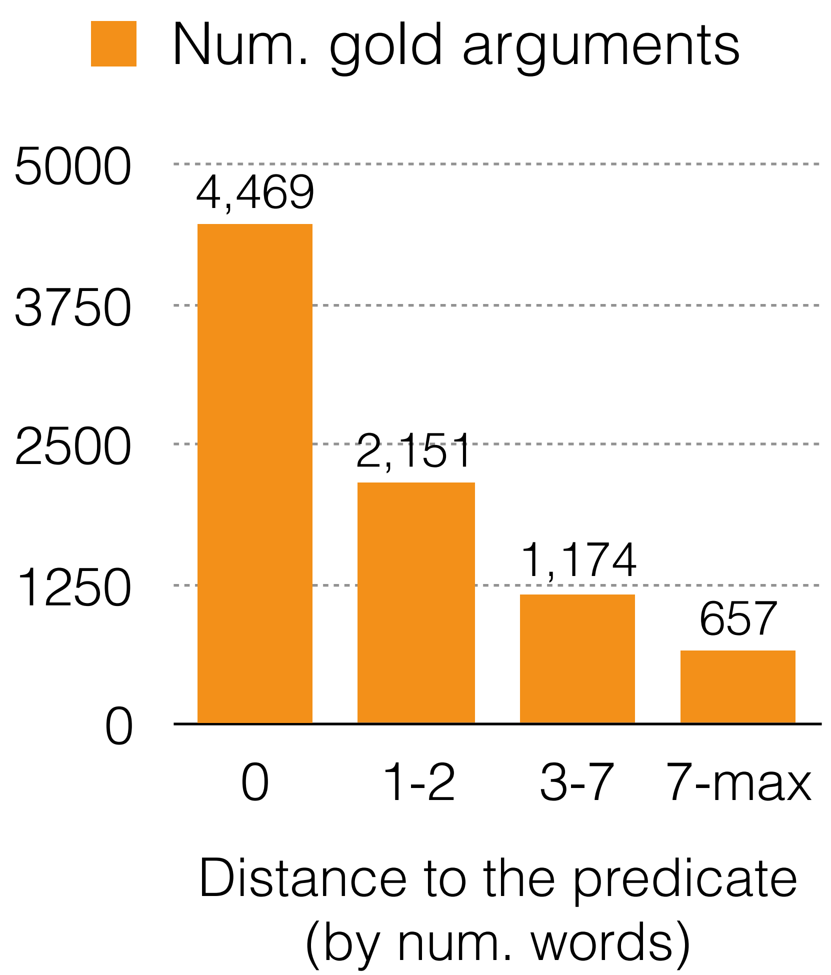


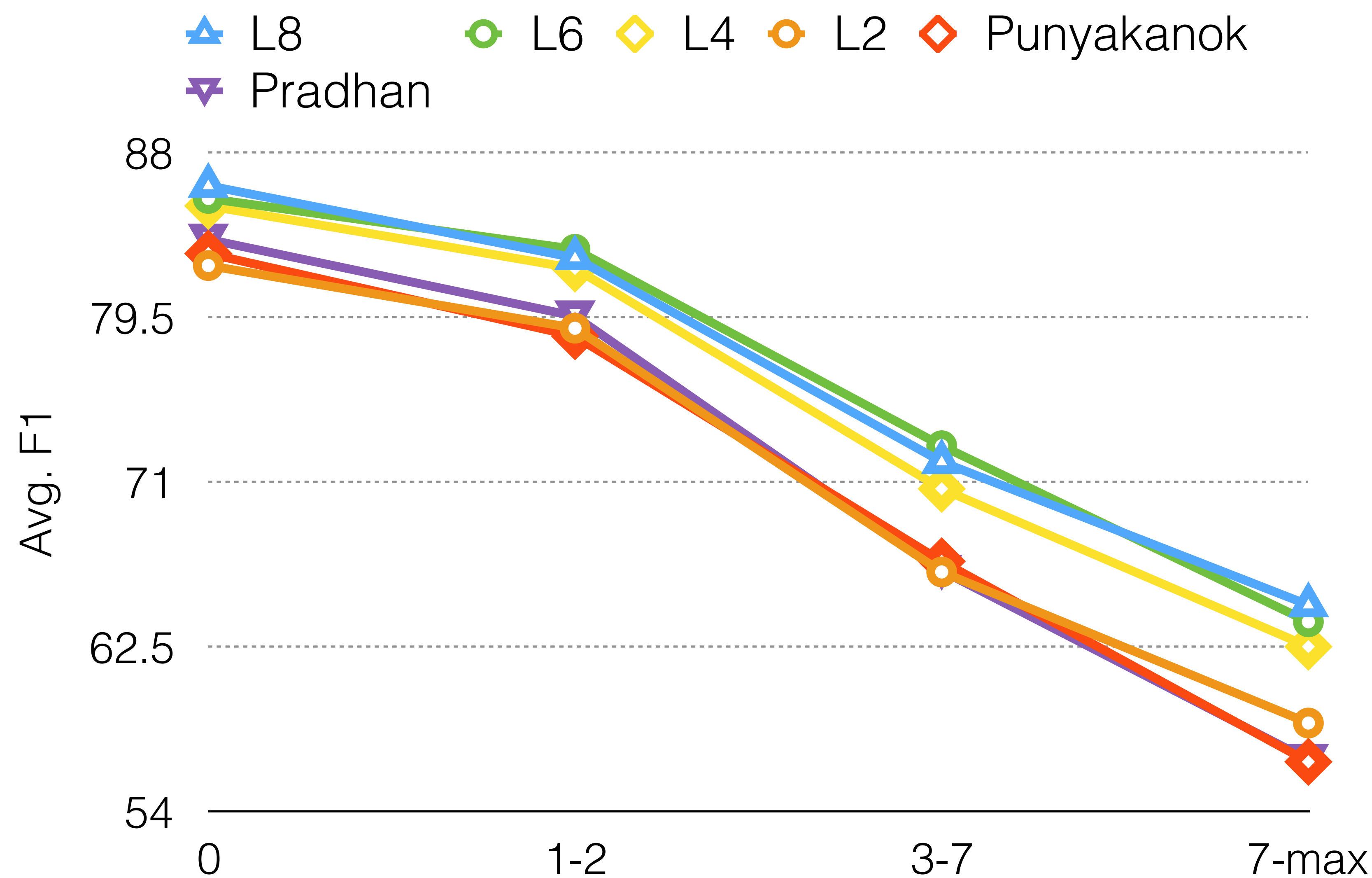
Long-range Dependencies



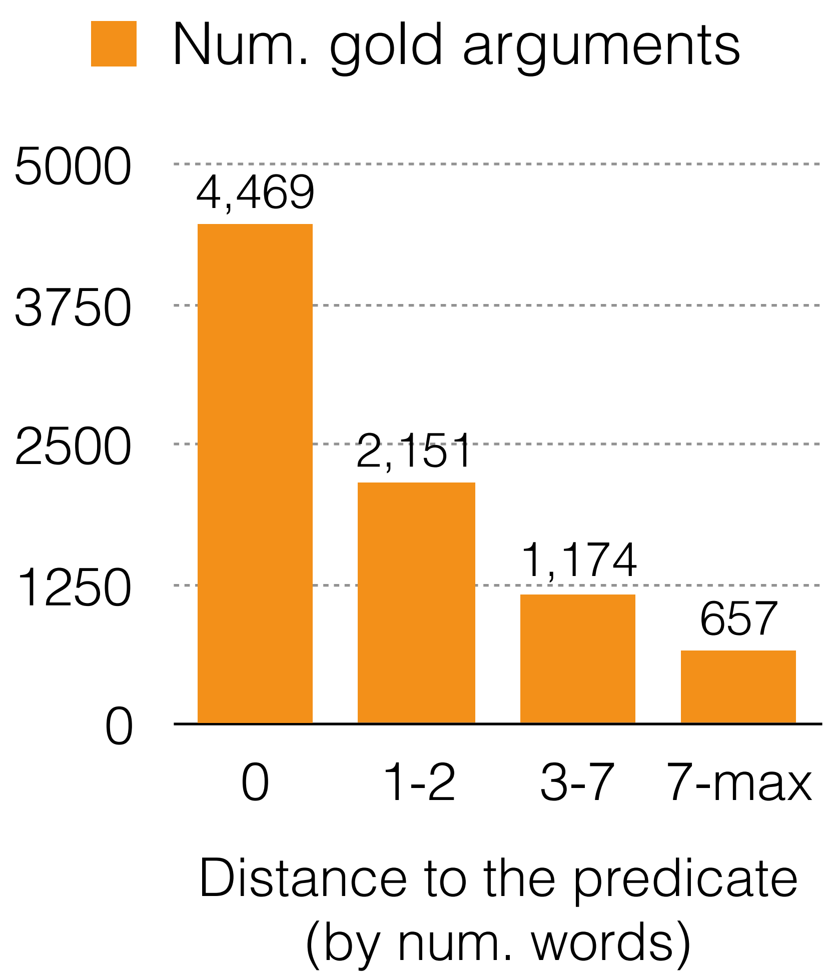


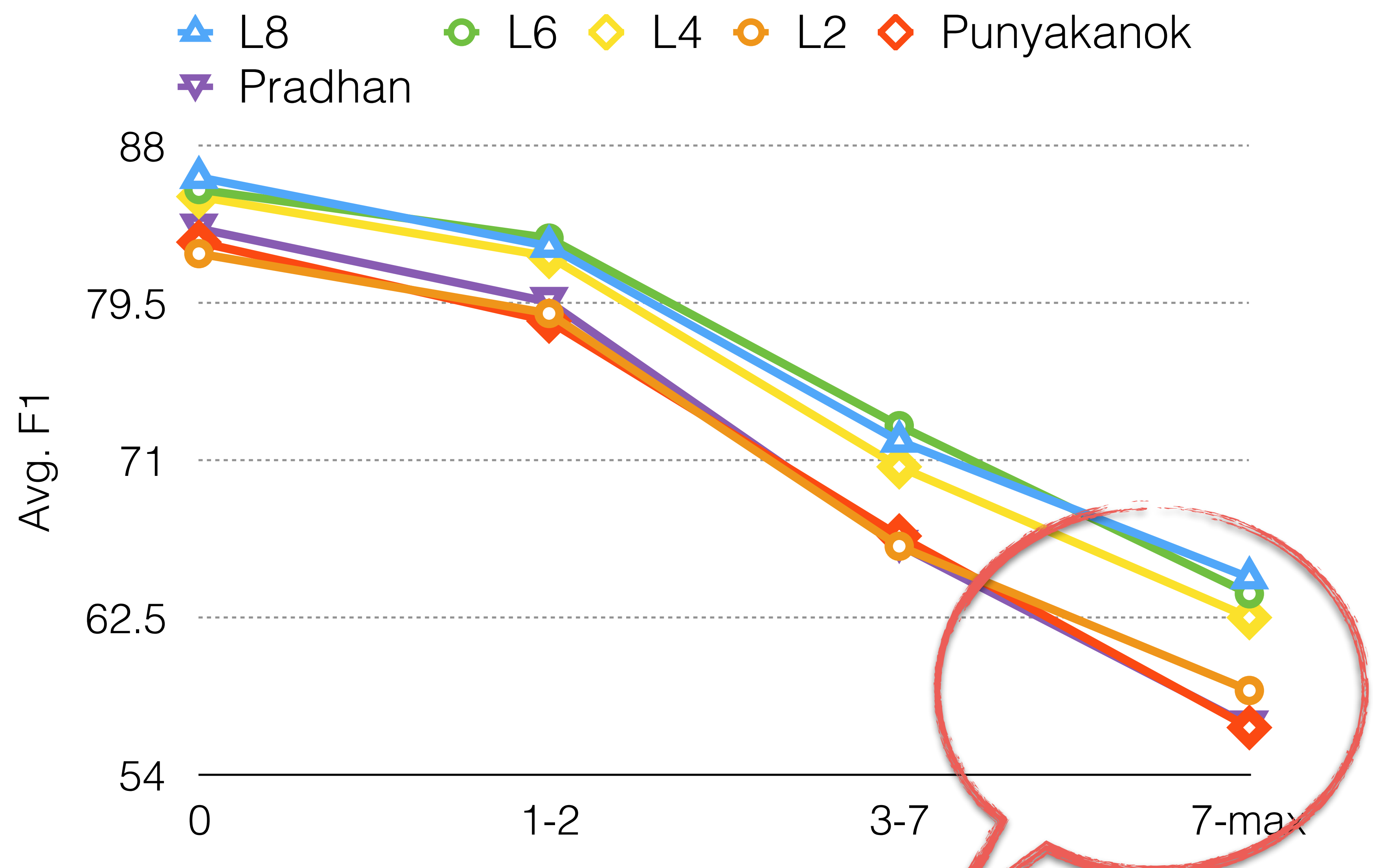
Long-range Dependencies





Long-range Dependencies

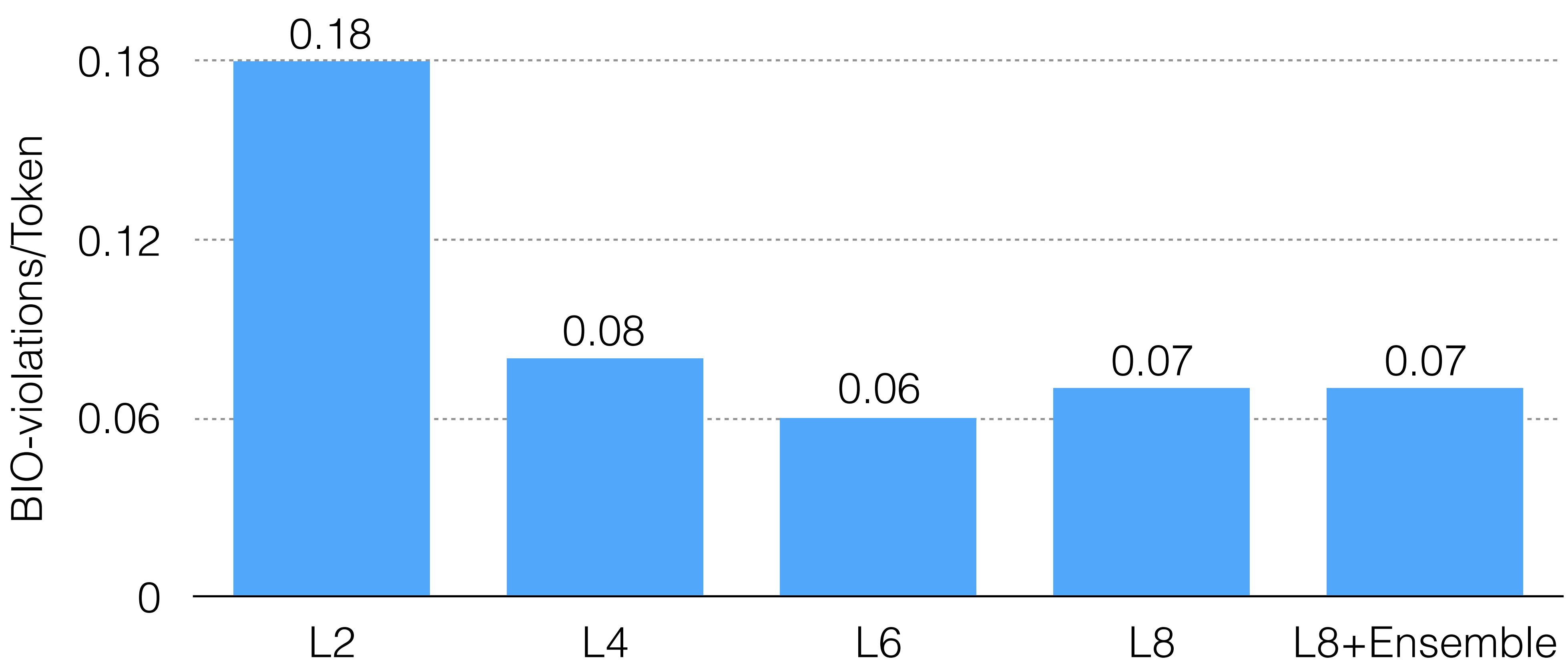




Deep models' performance deteriorates slower on long-distance predictions

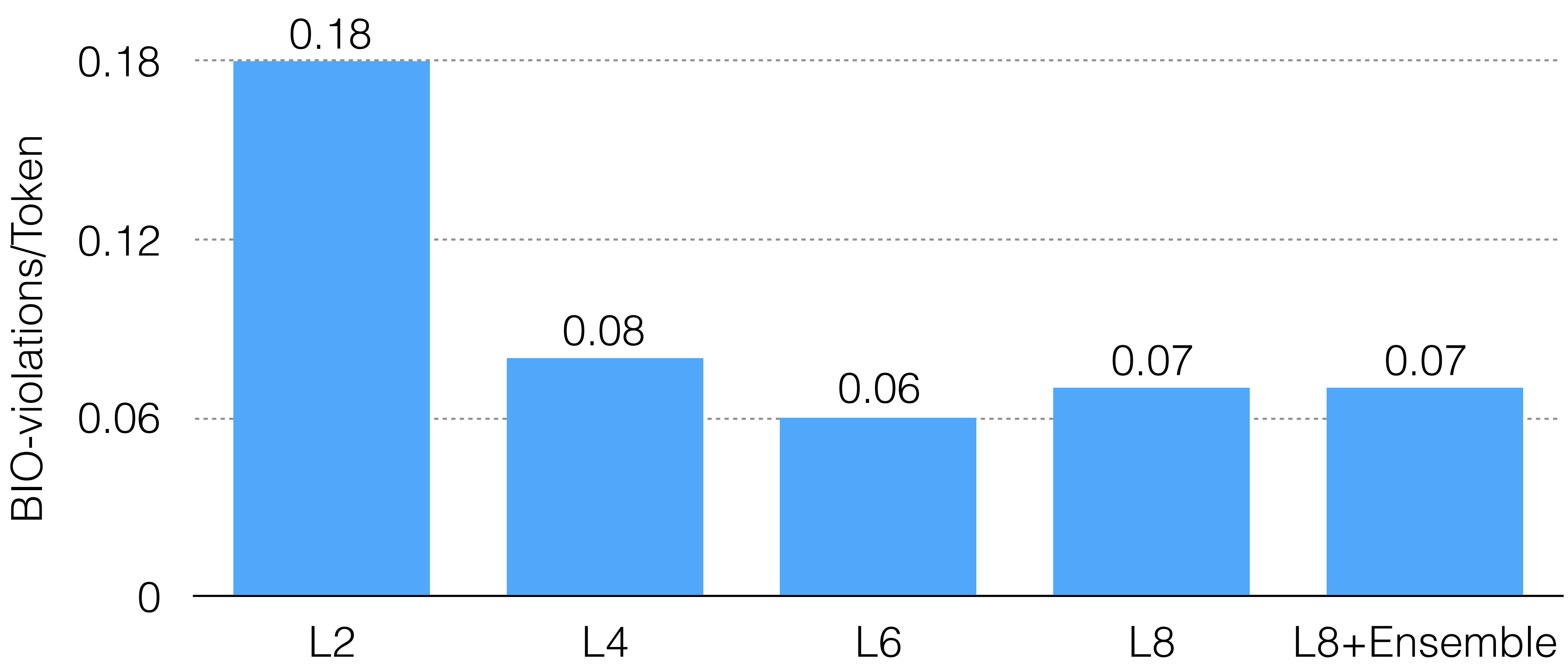
Structural Consistency

e.g. (B-ArgX, I-ArgY) or (O, I-ArgY)



Structural Consistency

e.g. (B-ArgX, I-ArgY) or (O, I-ArgY)



Deeper models (with 4+ layers) generate more consistent BIO sequences.

Question (3): Can syntax still help SRL?

Recap

— PropBank SRL is annotated on top of the PTB syntax.

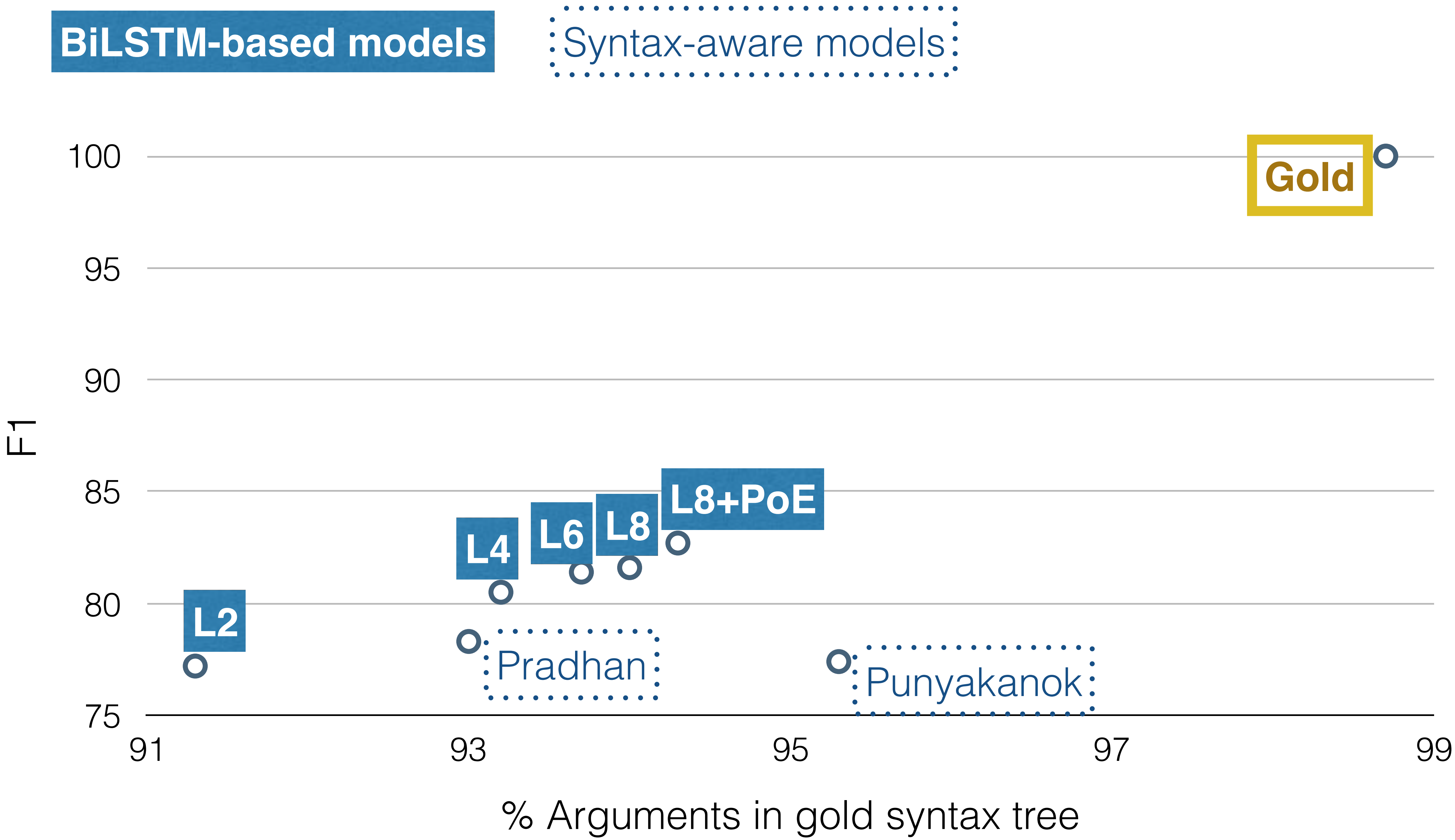
— More than 98% of the gold SRL spans are syntactic constituents.

Analysis

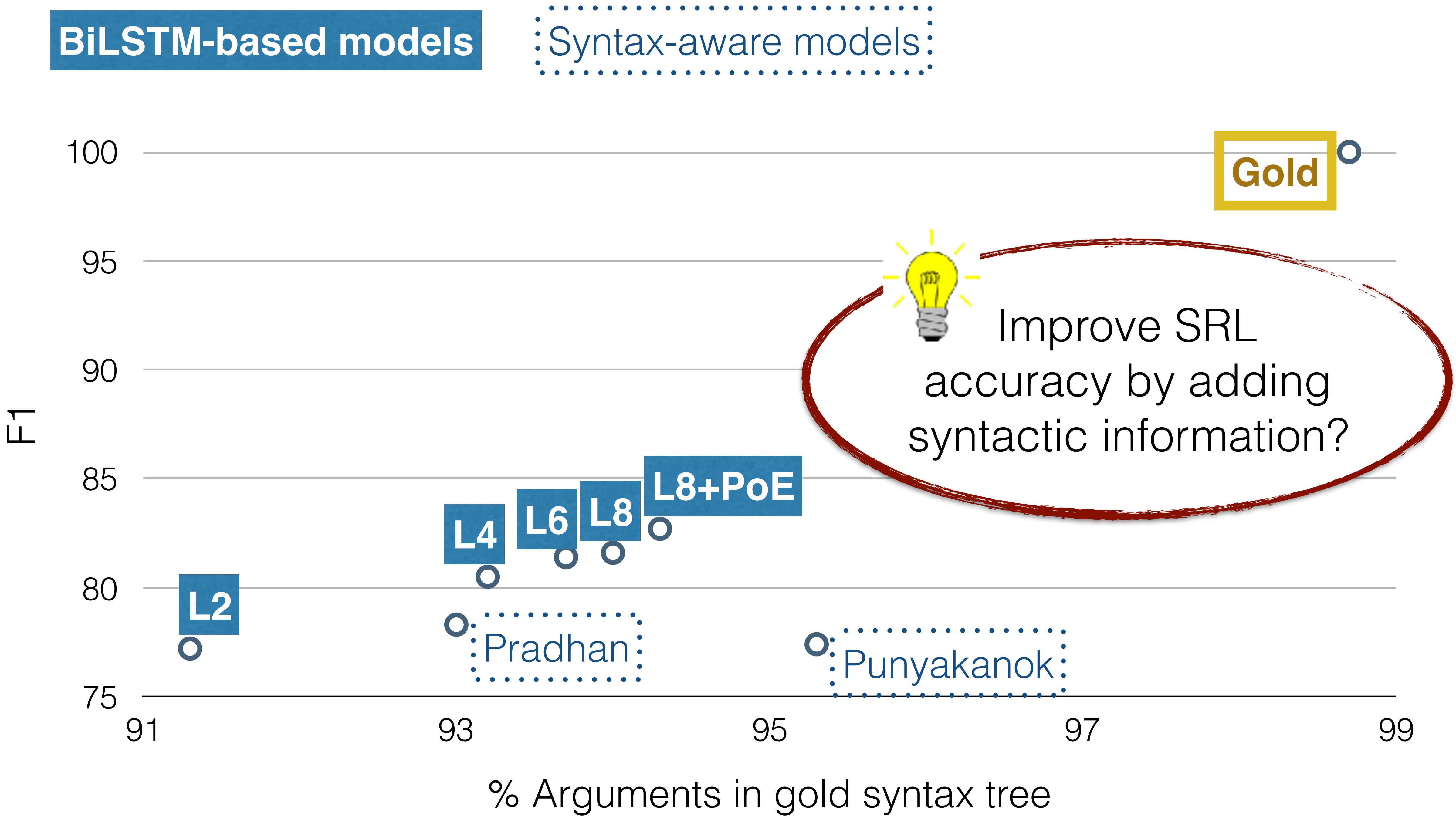
— At decoding time, make predicted argument spans agree with given syntactic structure.

— See if SRL performance increases.

Can Syntax Still Help?



Can Syntax Still Help?



Can Syntax Still Help?

Constrained Decoding with Syntax

[The cats] $\in \text{Syntax Tree}$

[hats and the dogs] $\notin \text{Syntax Tree}$

[The cats] *love* [hats and the dogs] love bananas.

ARG0 ARG1

Can Syntax Still Help?

Constrained Decoding with Syntax

[The cats] $\in$ Syntax Tree

[hats and the dogs] $\notin$ Syntax Tree

[The cats] *love* [hats and the dogs] love bananas.

ARG0 ARG1

Penalize sequence score

Can Syntax Still Help?

Constrained Decoding with Syntax

[The cats] $\in \text{Syntax Tree}$

[hats and the dogs] $\notin \text{Syntax Tree}$

[The cats] *love* [hats and the dogs] love bananas.

ARG0 ARG1

Penalize sequence score

Sequence score: $$\sum_{i=1}^{t} \log p(\text{tag}_t \mid \text{sentence}) - \mathcal{C} \times \sum_{\text{span}} \mathbf{1}(\text{span} \notin \text{Syntax Tree})$$

Penalty strength

Num. arguments disagree w\ syntax

Can Syntax Still Help?

Constrained Decoding with Syntax

[The cats] $\in \text{Syntax Tree}$

[hats and the dogs] $\notin \text{Syntax Tree}$

[The cats] *love* [hats and the dogs] love bananas.

ARG0 ARG1

Penalize sequence score

Sequence score: $\sum_{i=1}^{t} \log p(\text{tag}_t \mid \text{sentence}) - \mathcal{C} \times \sum_{\text{span}} \mathbf{1}(\text{span} \notin \text{Syntax Tree})$

Penalty strength

Num. arguments disagree w\ syntax

- Constraints are not locally decomposable.
- A* search (Lewis and Steedman 2014) for a sequence with highest score.

Can Syntax Still Help?

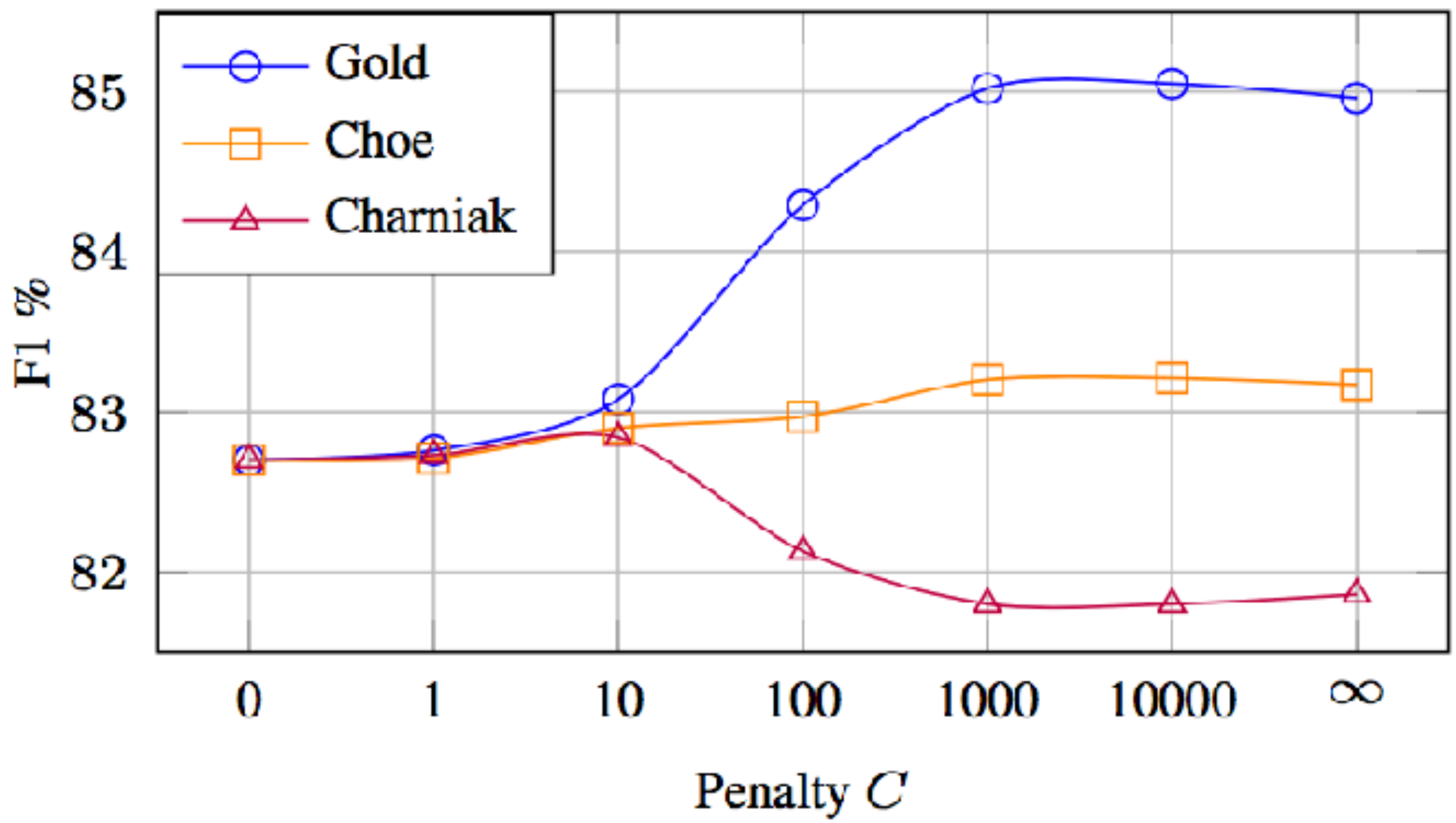
Constrained Decoding with Syntax

Charniak: A maximum-entropy-inspired parser, Charniak, 2000
Choe: Parsing as language modeling, Choe and Charniak, 2016
(State of the art)

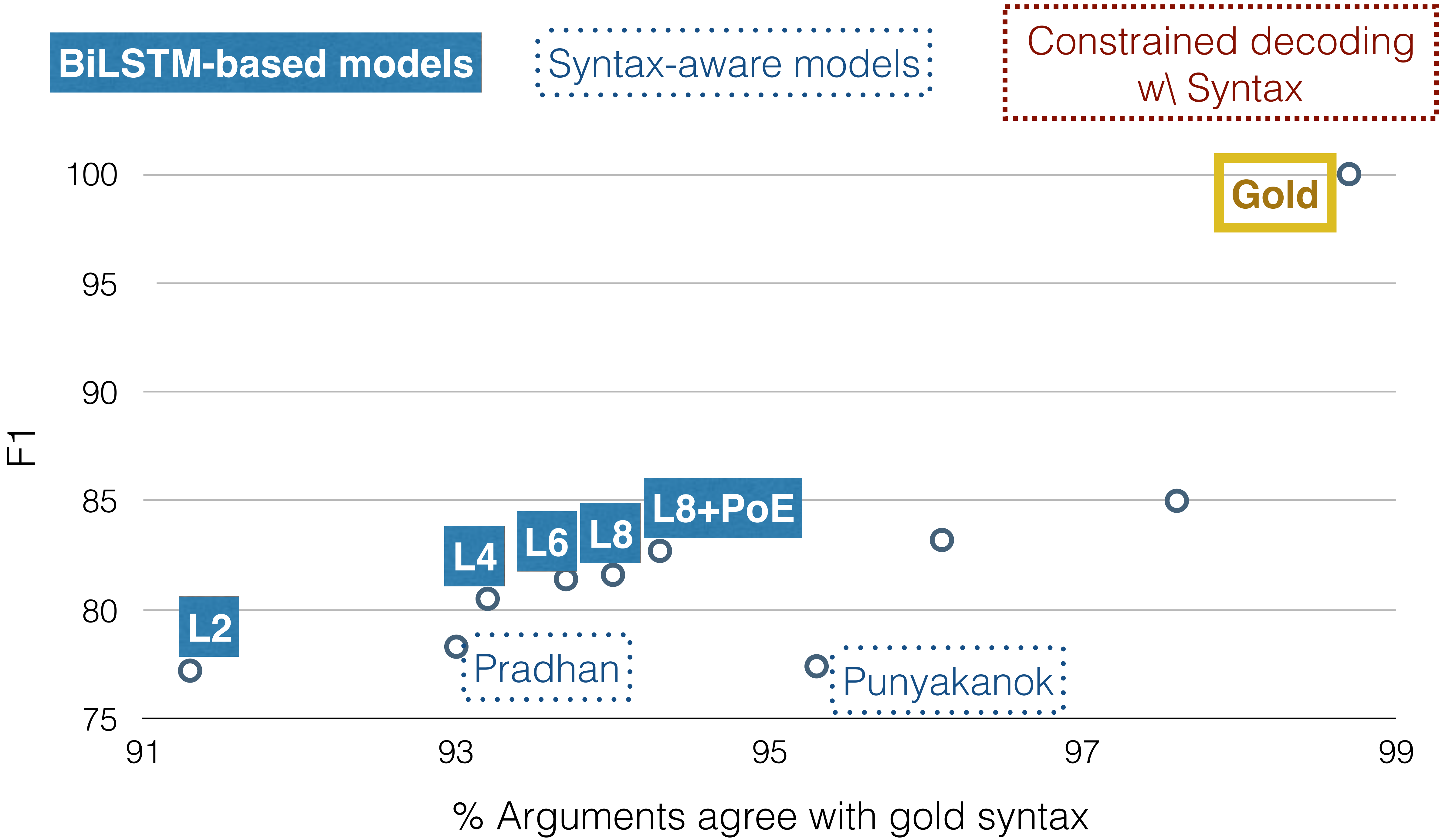
Can Syntax Still Help?

Constrained Decoding with Syntax

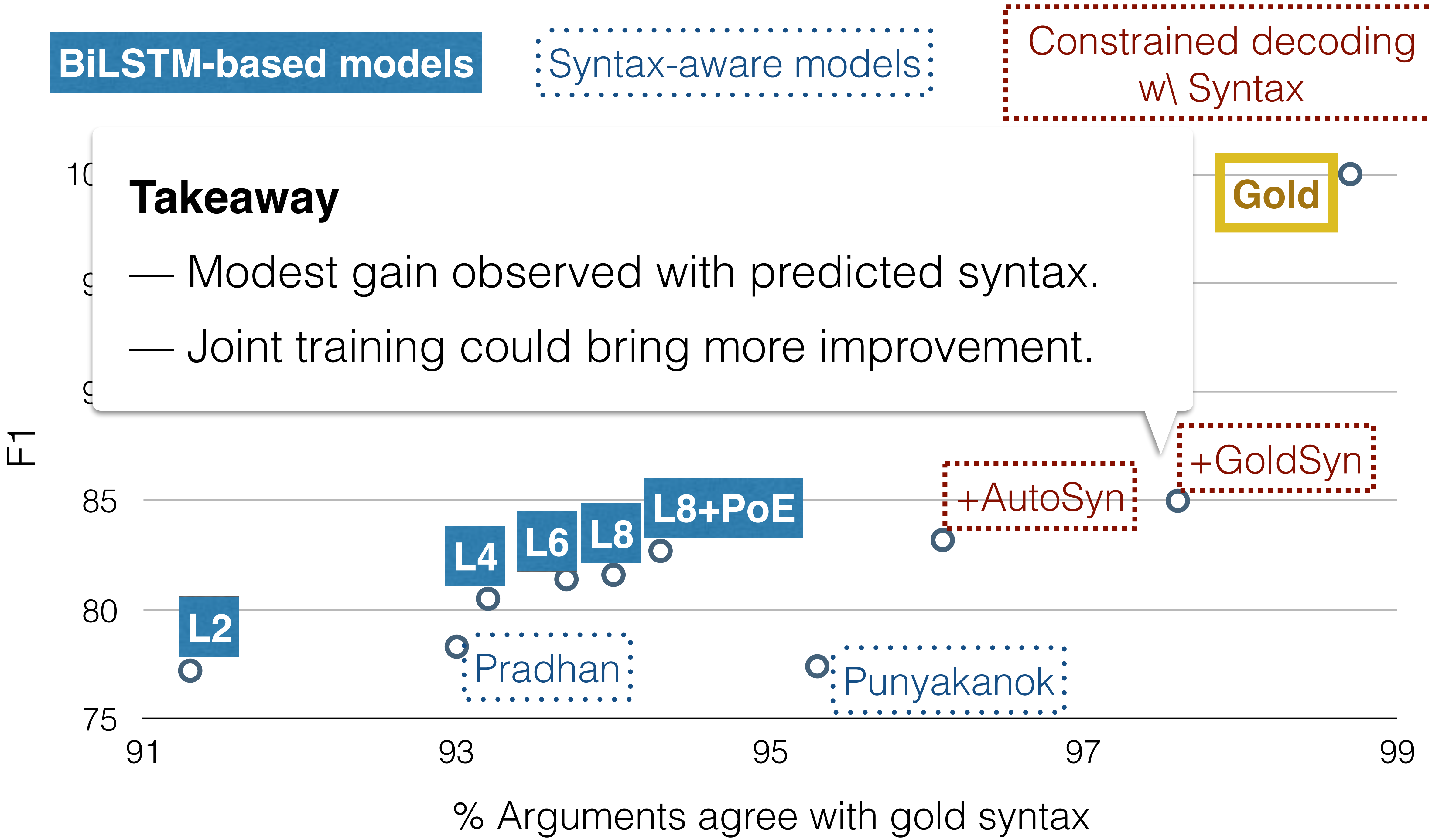
Charniak: A maximum-entropy-inspired parser, Charniak, 2000
Choe: Parsing as language modeling, Choe and Charniak, 2016 (State of the art)



Can Syntax Still Help?



Can Syntax Still Help?

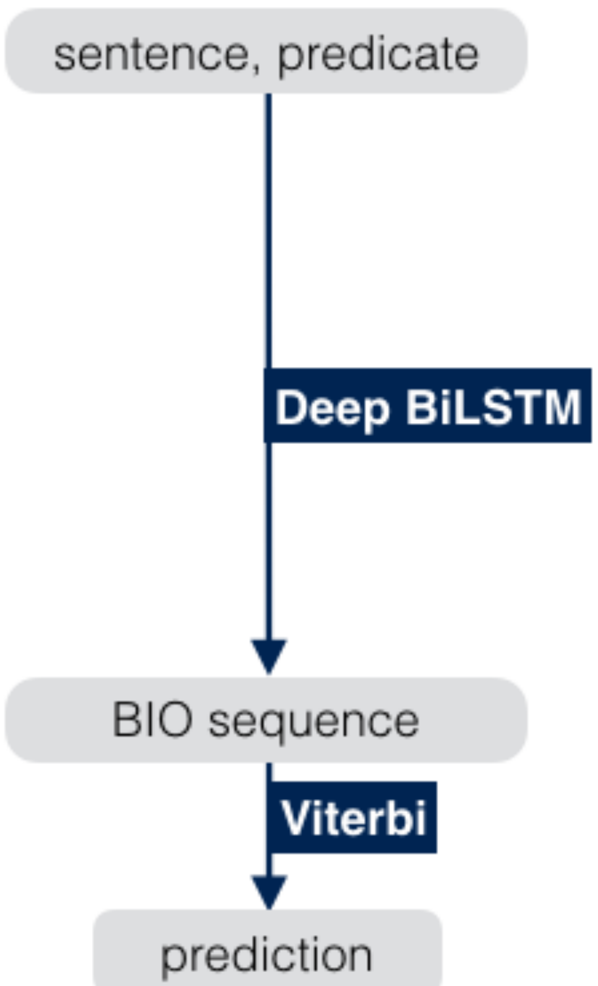


Takeaway

— Modest gain observed with predicted syntax.

— Joint training could bring more improvement.

Contributions (Neural SRL)



- New state-of-the-art deep network for end-to-end SRL.
- Code and models will be publicly available at: https://github.com/luheng/deep_srl
- In-depth error analysis indicating where the models work well and where they still struggle.
- Syntax-based experiments pointing towards directions for future improvements.

Long-term Plan for Improving SRL

Step 1: Collect more data for SRL

— Question-Answer Driven Semantic Role Labeling (QA-SRL)

— Human-in-the-Loop Parsing

✓

Step 2: Build accurate SRL model

— Neural Semantic Role Labeling (for PropBank SRL)

Step 3: SRL system for many domains

— *Future work …*

Long-term Plan for Improving SRL

Step 1: Collect more data for SRL

— Question-Answer Driven Semantic Role Labeling (QA-SRL)

— Human-in-the-Loop Parsing

✓

Step 2: Build accurate SRL model

— Neural Semantic Role Labeling (for PropBank SRL)

✓

Step 3: SRL system for many domains

— *Future work …*

Thanks!

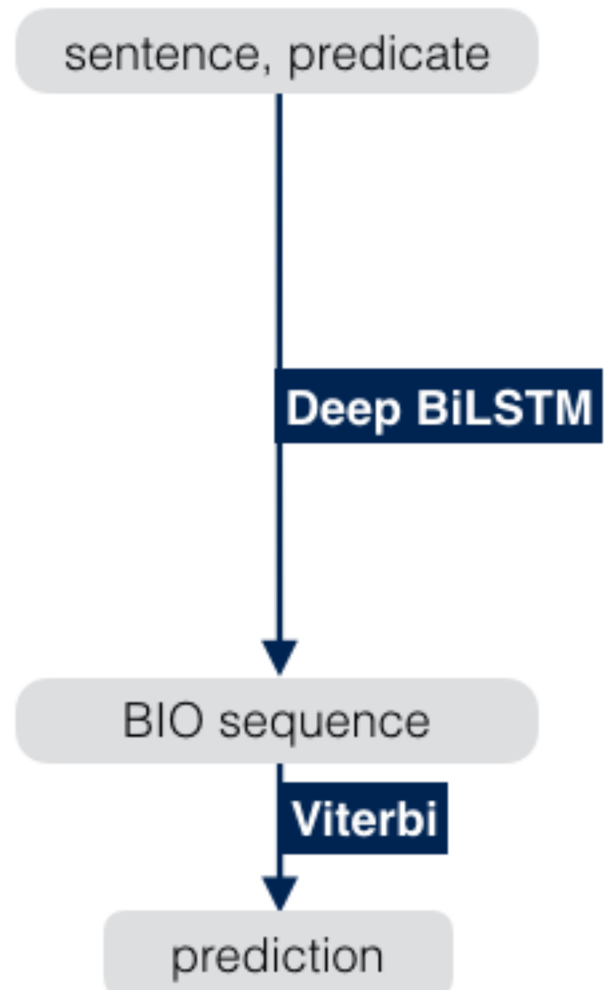
Code will be available at: https://github.com/luheng/deep_srl

Problem

Frame: *break.01*

role	description
ARG0	breaker
ARG1	thing broken
ARG2	instrument
ARG3	pieces
ARG4	broken away from what?

Model



Analysis

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	76	13	6	14	2	0	0	0	0	0
A1	16	74	25	0	0	18	9	11	19	2
A2	2	5	31	52	10	45	26	46	19	0
A3	1	0	1	57	2	0	0	0	19	2
ADV	0	0	0	5	33	0	11	33	19	5
DIR	0	0	3	5	0	27	9	2	0	0
LOC	1	2	7	0	2	0	34	11	0	2
MNR	1	0	7	29	21	0	0	43	0	3
PNC	0	1	3	5	0	9	3	2	44	0
TMP	0	2	3	0	26	9	20	7	0	71