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

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)

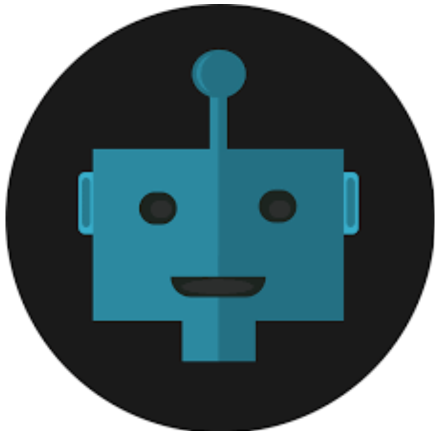
role label

who
what
when
where
why
…

predicate argument

Applications

Question Answering



Information Extraction

Machine Translation

Semantic Role Labeling (SRL) - Example

The robot *broke* my mug with a wrench.

My mug *broke* into pieces.

Semantic Role Labeling (SRL) - Example

v **obj**

The robot *broke* my mug with a wrench.

breaker thing broken instrument

subj **v**

My mug *broke* into pieces.

thing broken pieces (final state)

Semantic Role Labeling (SRL) - Example

v obj

The robot *broke* my mug with a wrench.

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

subj v

My mug *broke* into pieces.

thing broken — ARG1 | pieces (final state) — ARG3

Frame: *break.01*

role	description
ARG0	breaker
ARG1	thing broken
ARG2	instrument
ARG3	pieces
...	...

The Proposition Bank (PropBank)

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

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
…	

The Proposition Bank (PropBank)

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

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

Frame: *break.01*

role	description
ARG0	breaker
ARG1	thing broken
ARG2	instrument

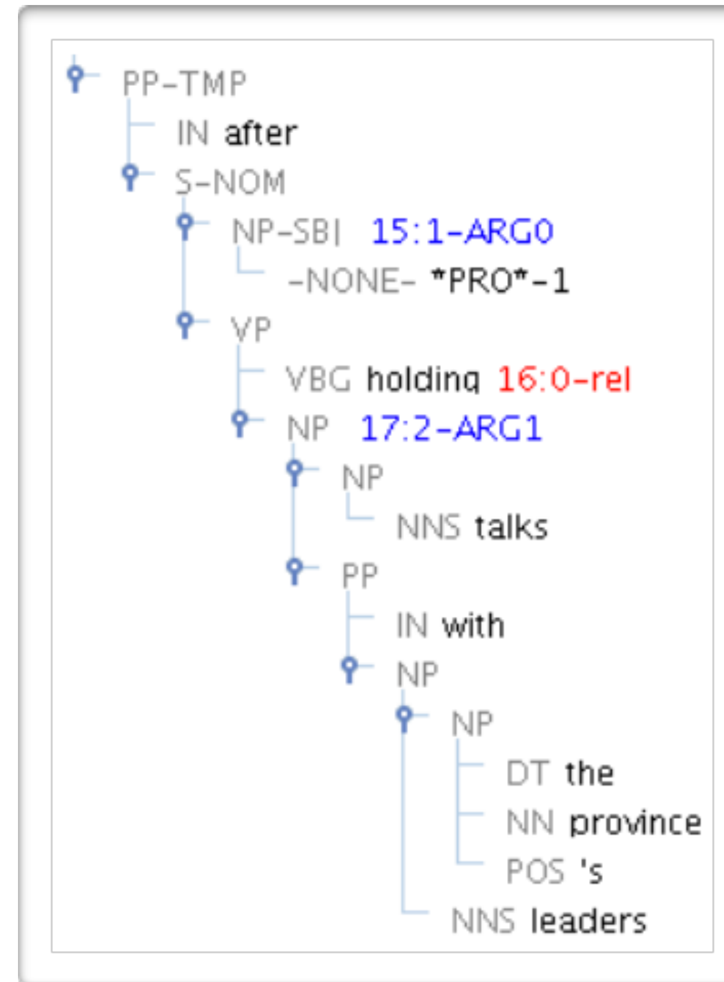
Frame: *buy.01*

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
…	

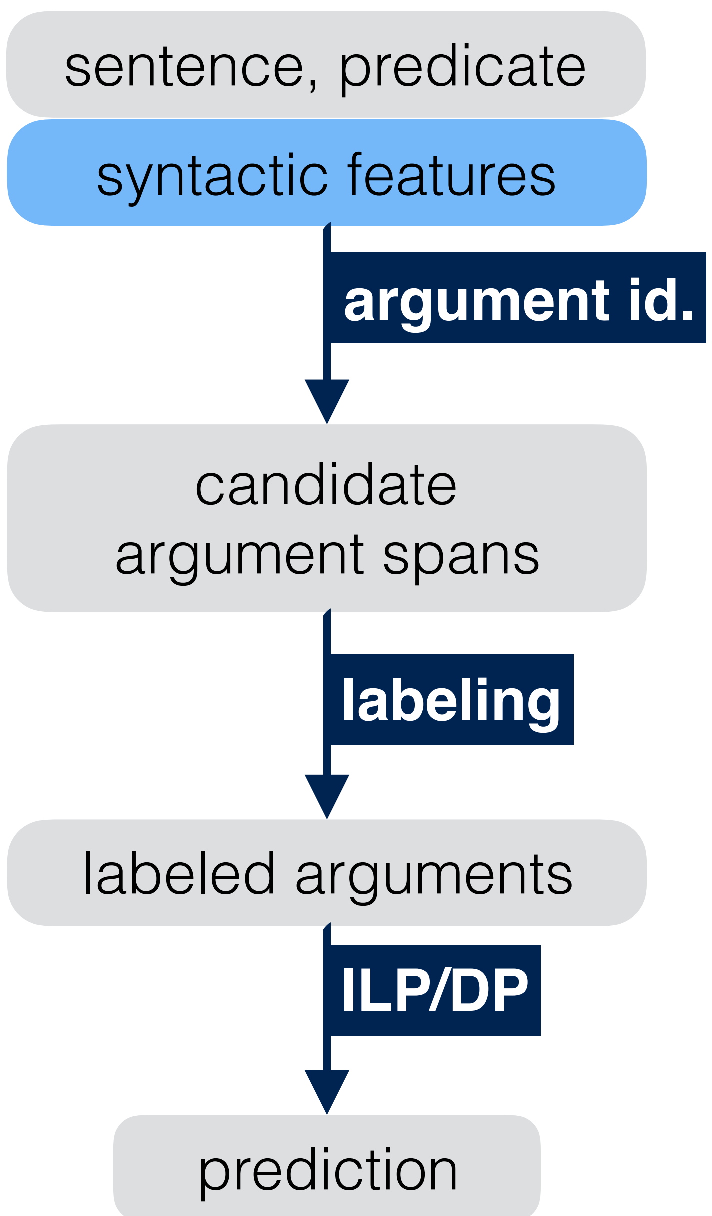
Annotated on top of the Penn Treebank Syntax



PropBank Annotation Guidelines, Bonial et al., 2010

SRL Systems

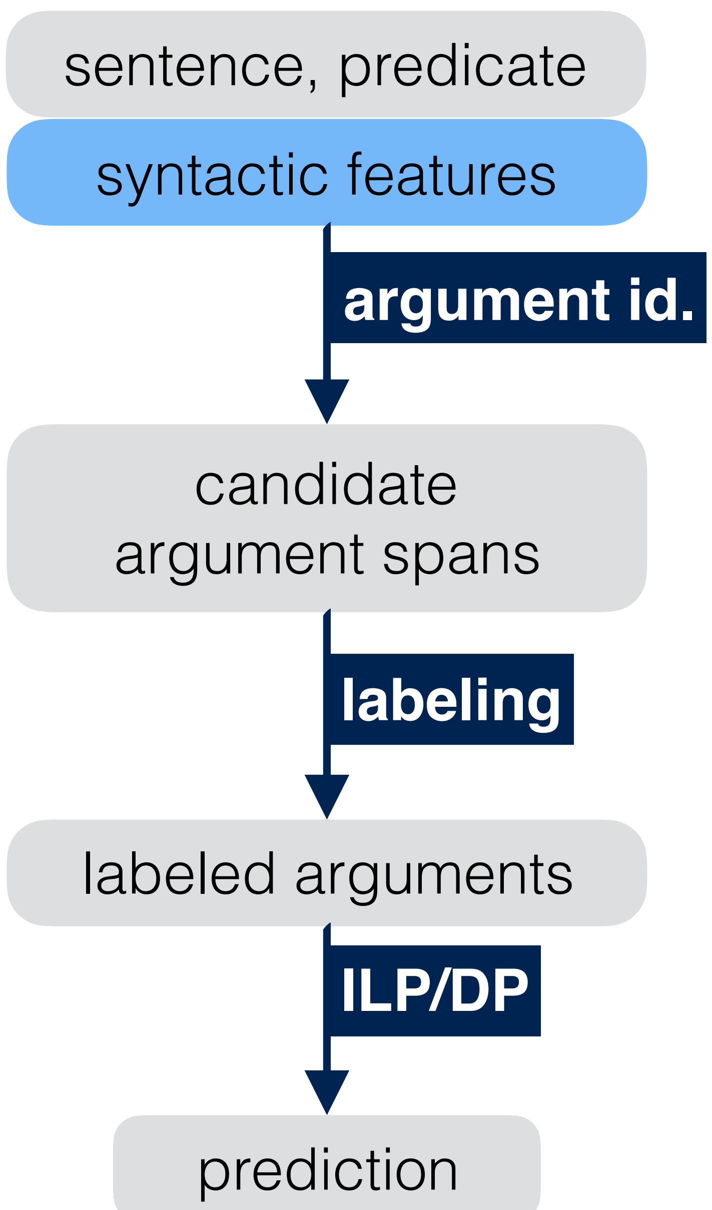
Pipeline Systems



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

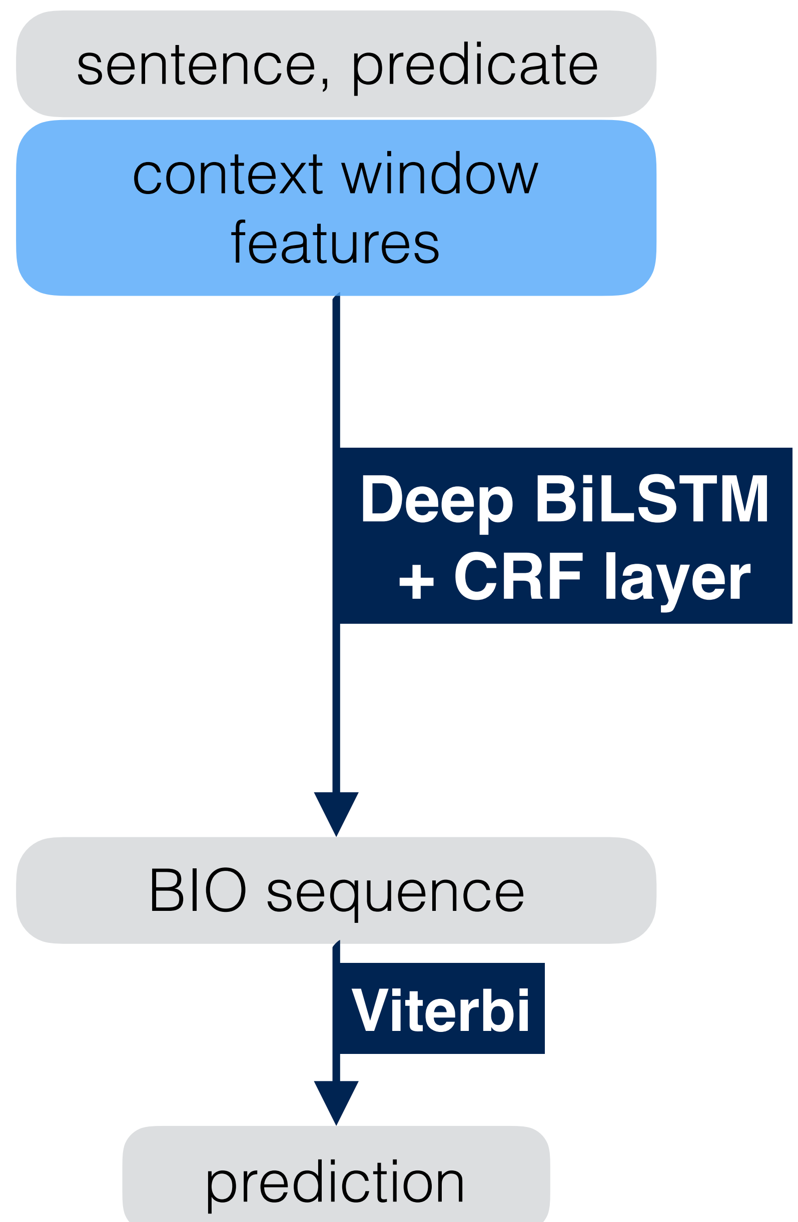
SRL Systems

Pipeline Systems



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

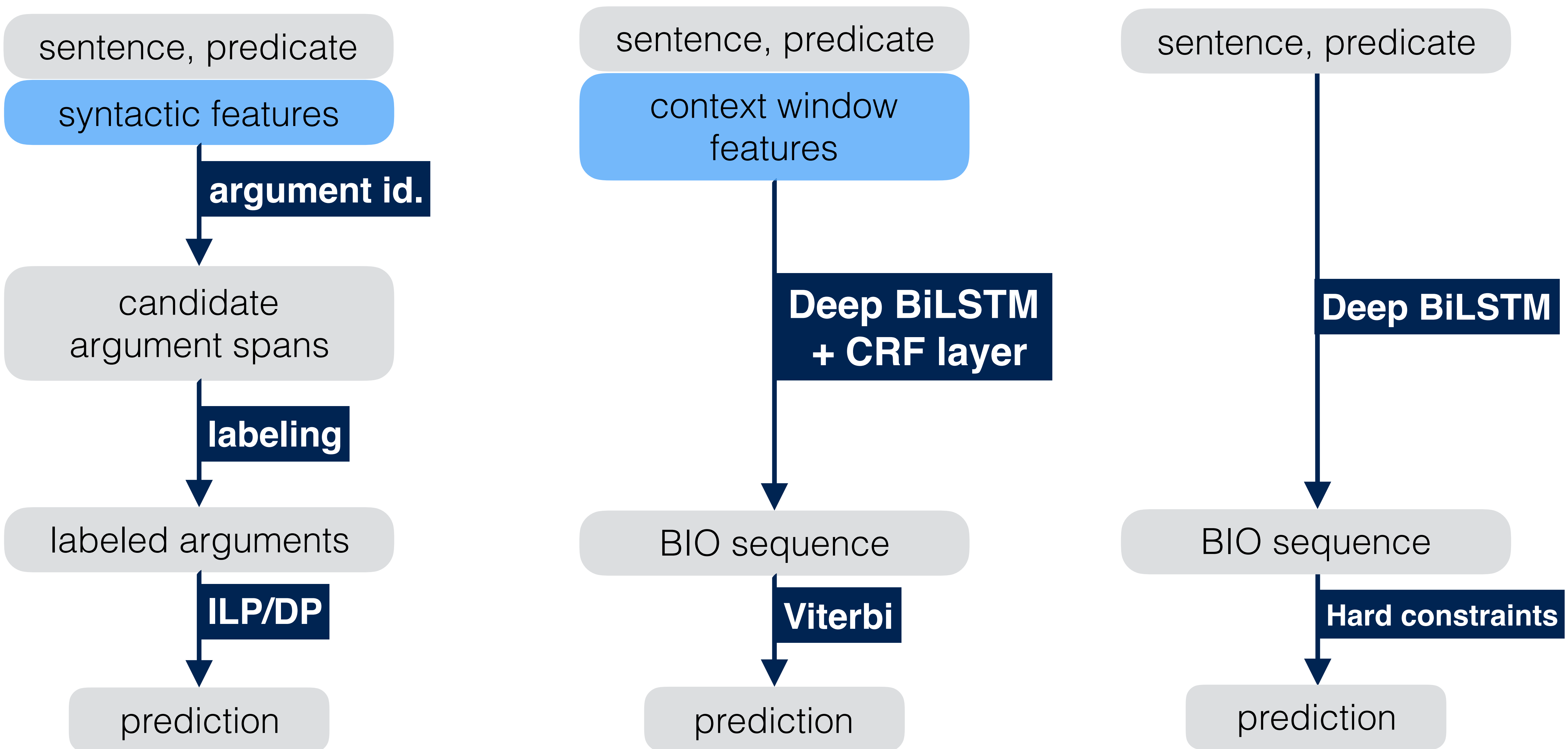
End-to-end Systems



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

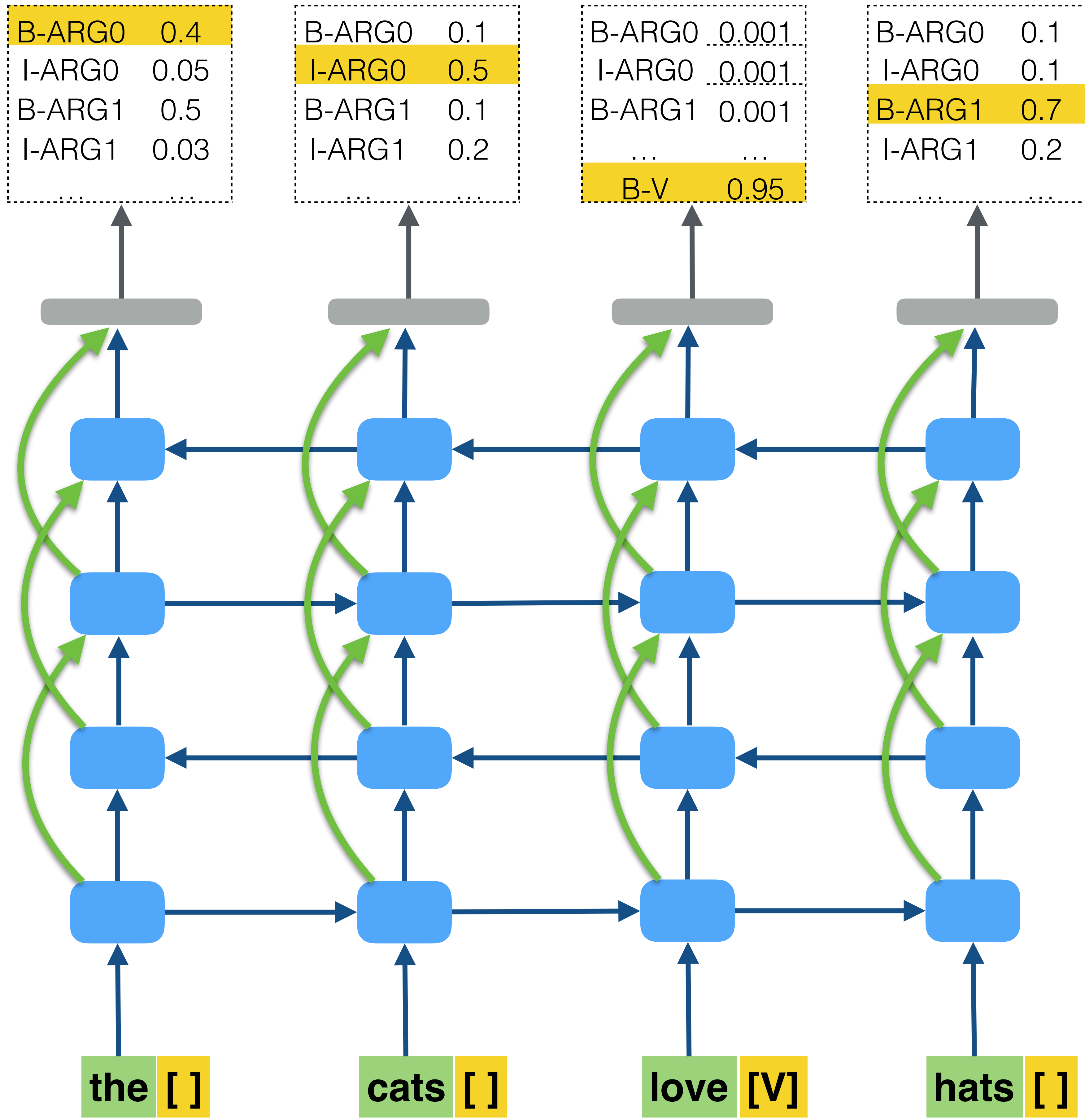
SRL Systems

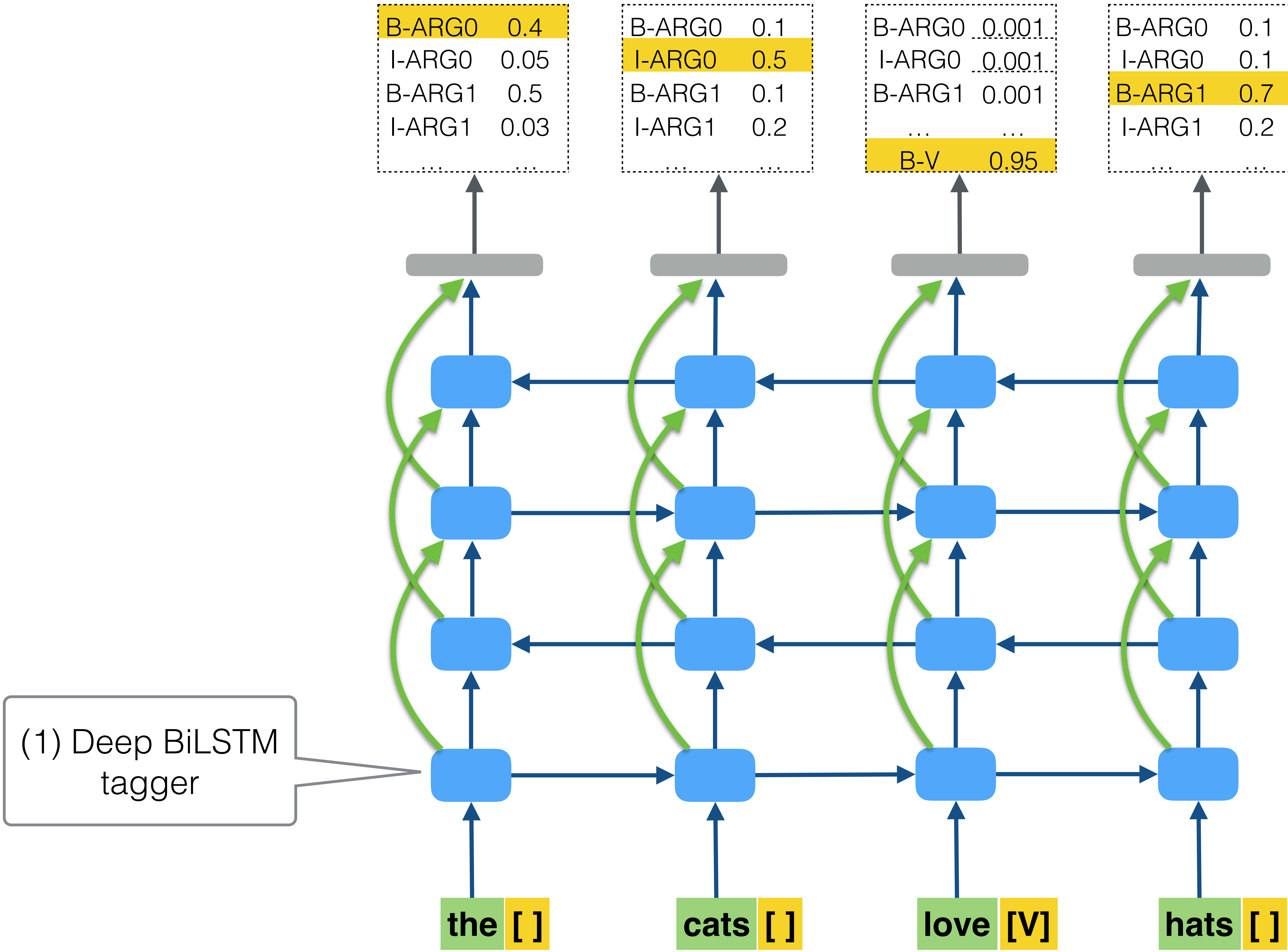
Pipeline Systems | End-to-end Systems | ***This work**

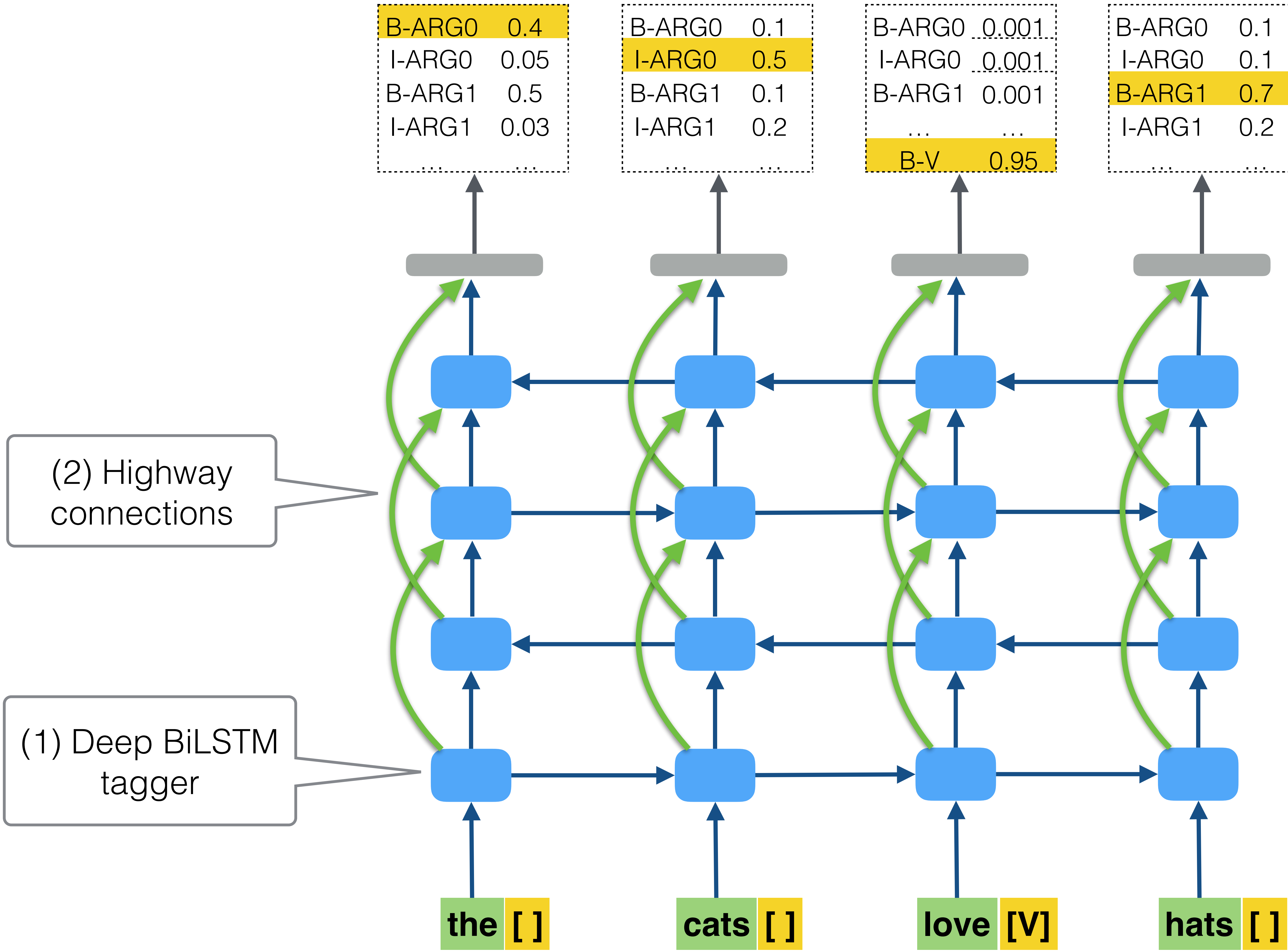


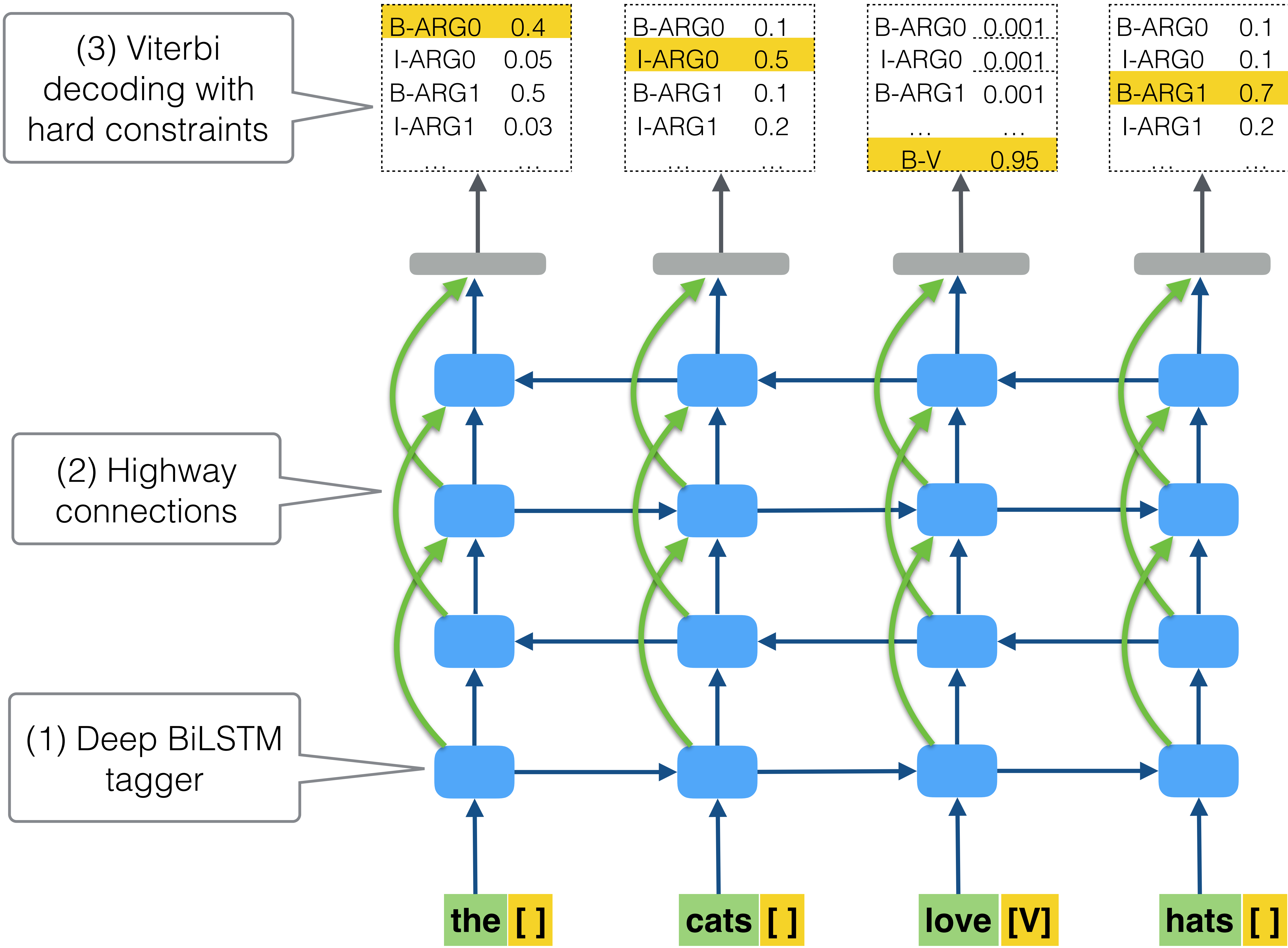
Punyakanok et al., 2008
Täckström et al., 2015
FitzGerald et al., 2015

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


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 [ ]

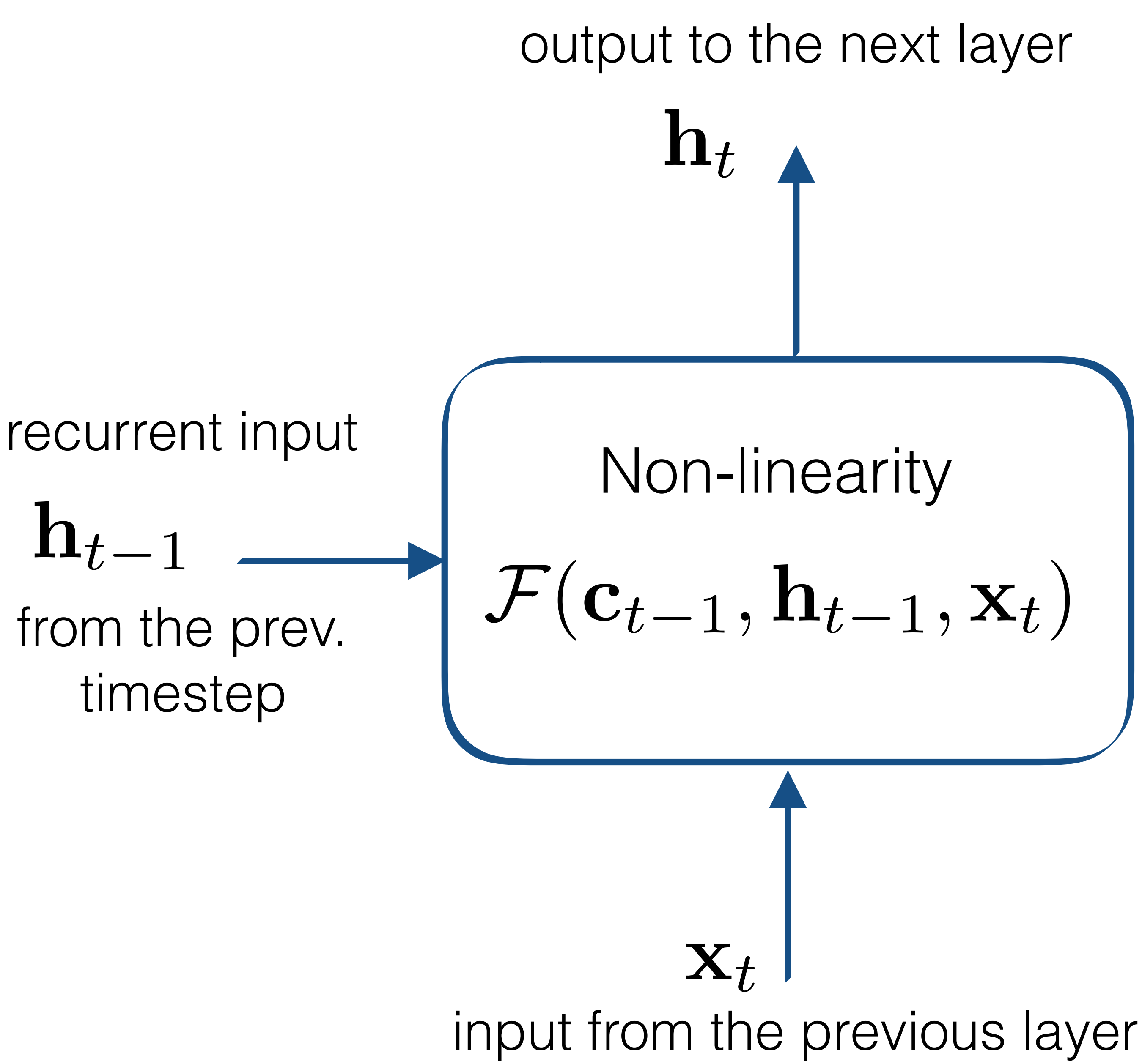

B-ARG0 0.4
I-ARG0 0.05
B-ARG1 0.5
I-ARG1 0.03
... ...
B-ARG0 0.1
I-ARG0 0.5
B-ARG1 0.1
I-ARG1 0.2
... ...
B-ARG0 0.001
I-ARG0 0.001
B-ARG1 0.001
... ...
B-V 0.95
B-ARG0 0.1
I-ARG0 0.1
B-ARG1 0.7
I-ARG1 0.2
... ...
(1) Deep BiLSTM tagger
the [ ]
cats [ ]
love [V]
hats [ ]


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 [ ]



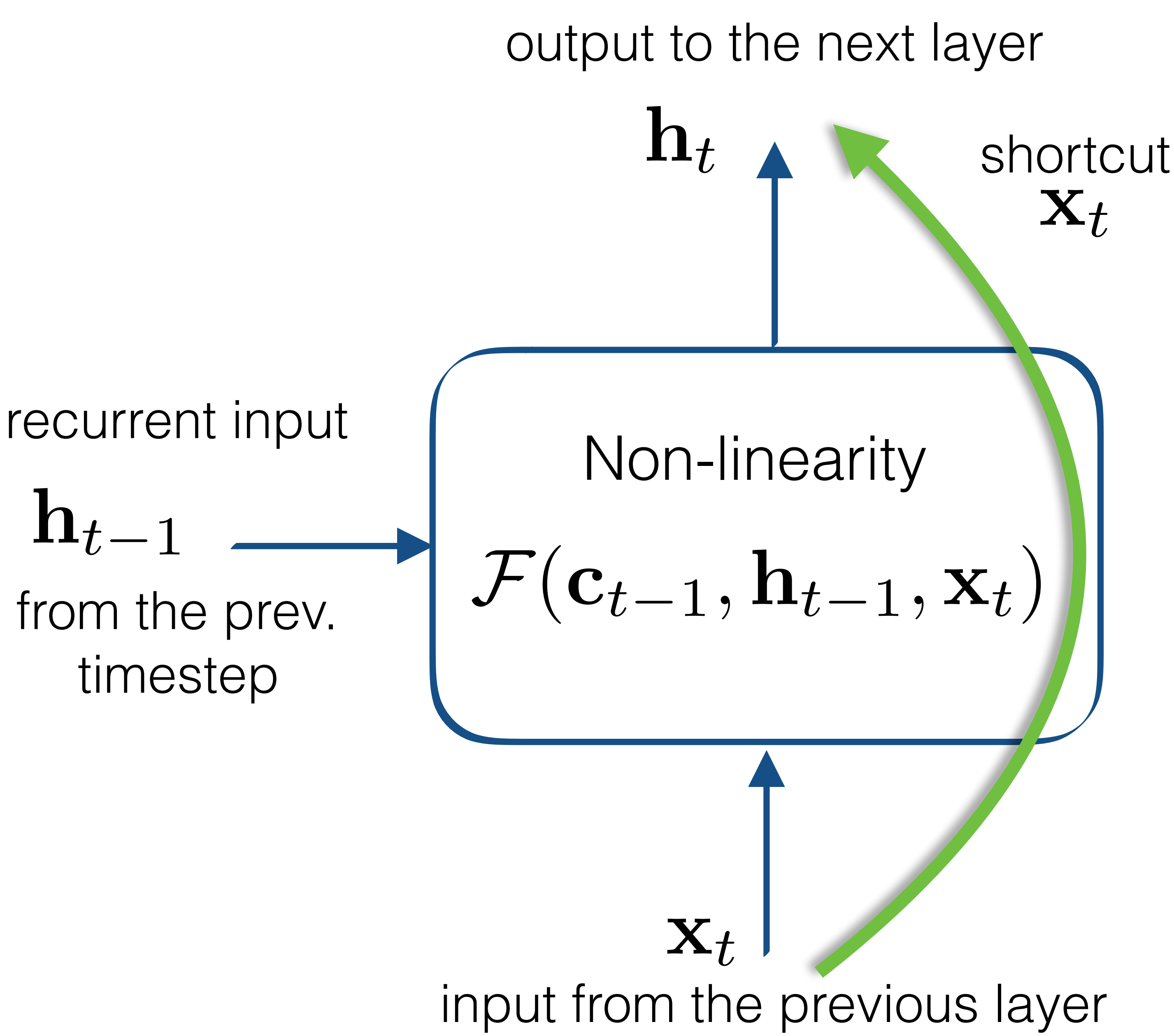
(3) Viterbi decoding with hard constraints
(2) Highway connections
(1) Deep BiLSTM tagger
B-ARG0 0.4
I-ARG0 0.05
B-ARG1 0.5
I-ARG1 0.03
B-ARG0 0.1
I-ARG0 0.5
B-ARG1 0.1
I-ARG1 0.2
B-ARG0 0.001
I-ARG0 0.001
B-ARG1 0.001
B-V 0.95
B-ARG0 0.1
I-ARG0 0.1
B-ARG1 0.7
I-ARG1 0.2
the [ ]
cats [ ]
love [V]
hats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

Model - Highway Connections



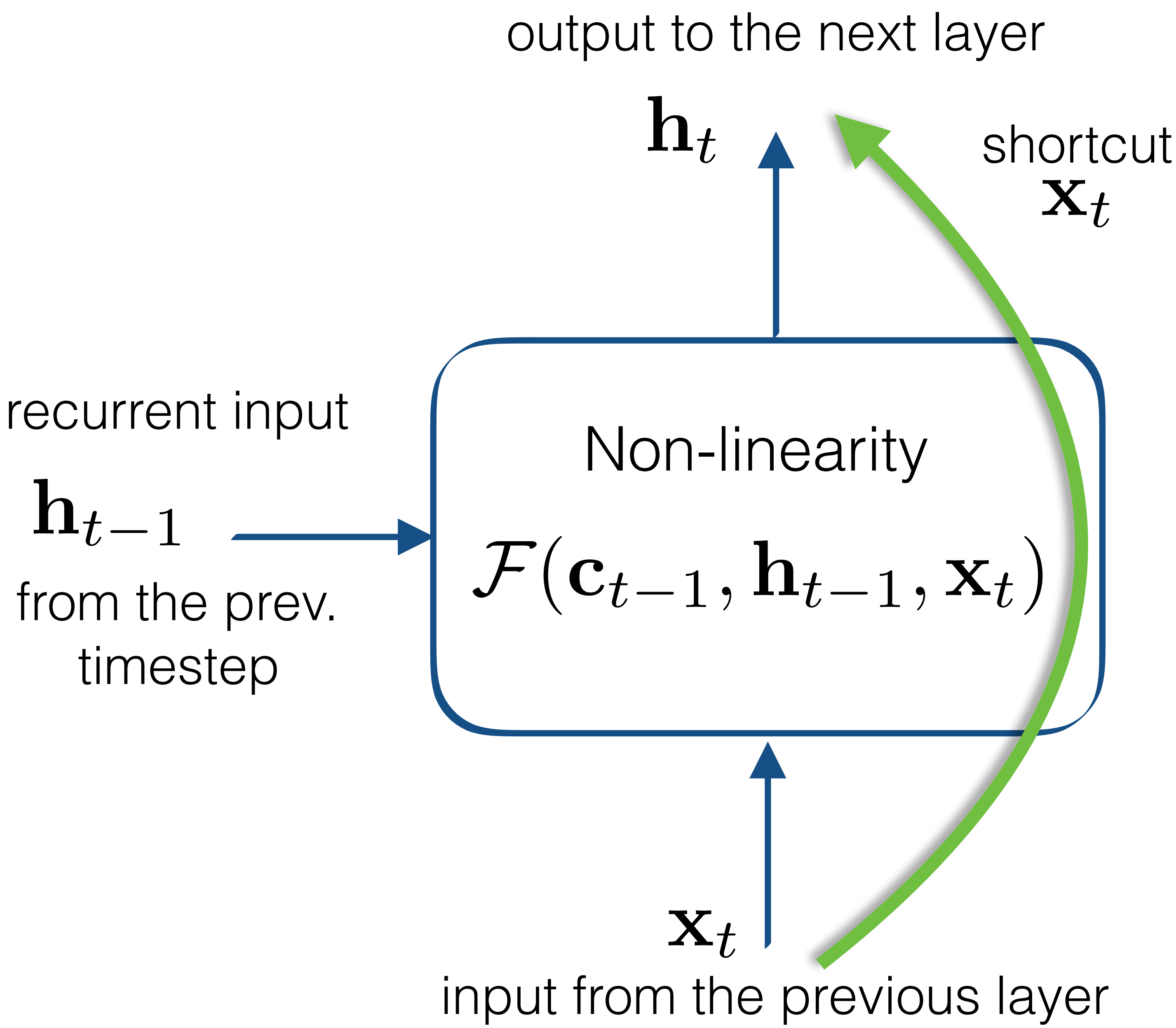
References:
Deep Residual Networks, Kaiming He, ICML 2016 Tutorial
Training Very Deep Networks, Srivastava et al., 2015

Model - Highway Connections



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

Model - Highway Connections



new output:

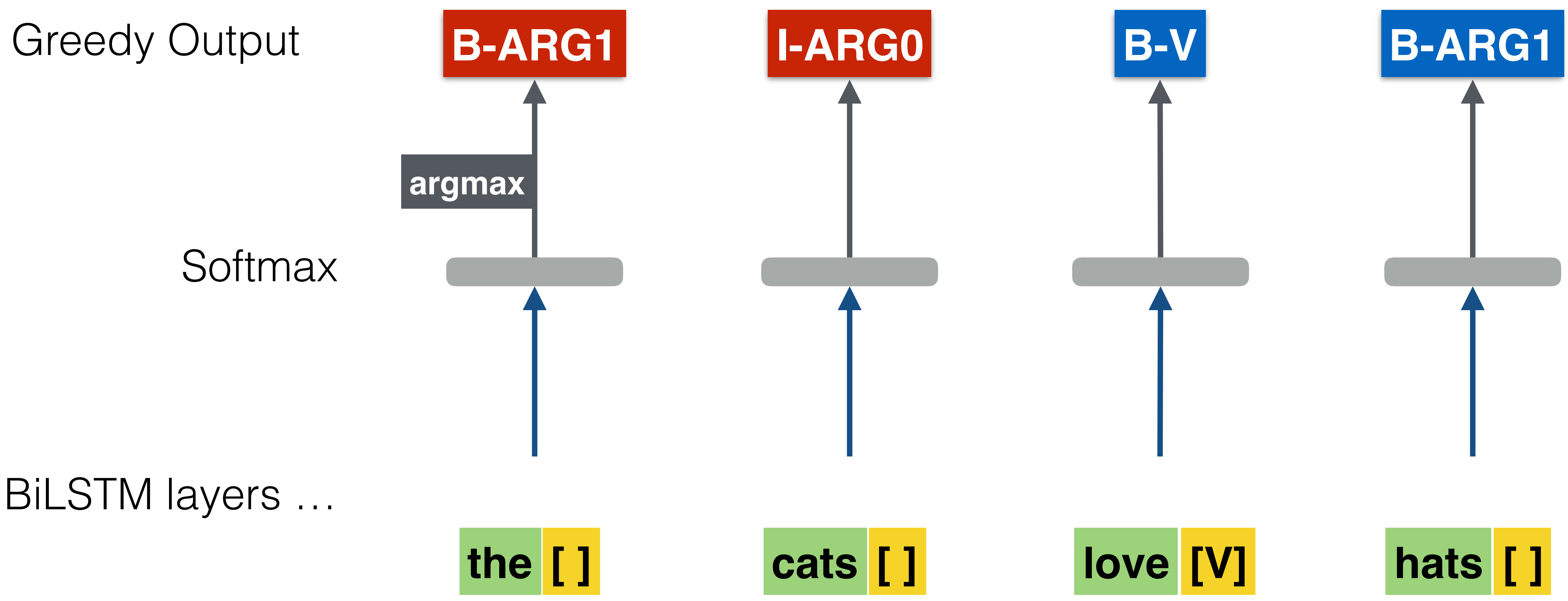
gated highway network:

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

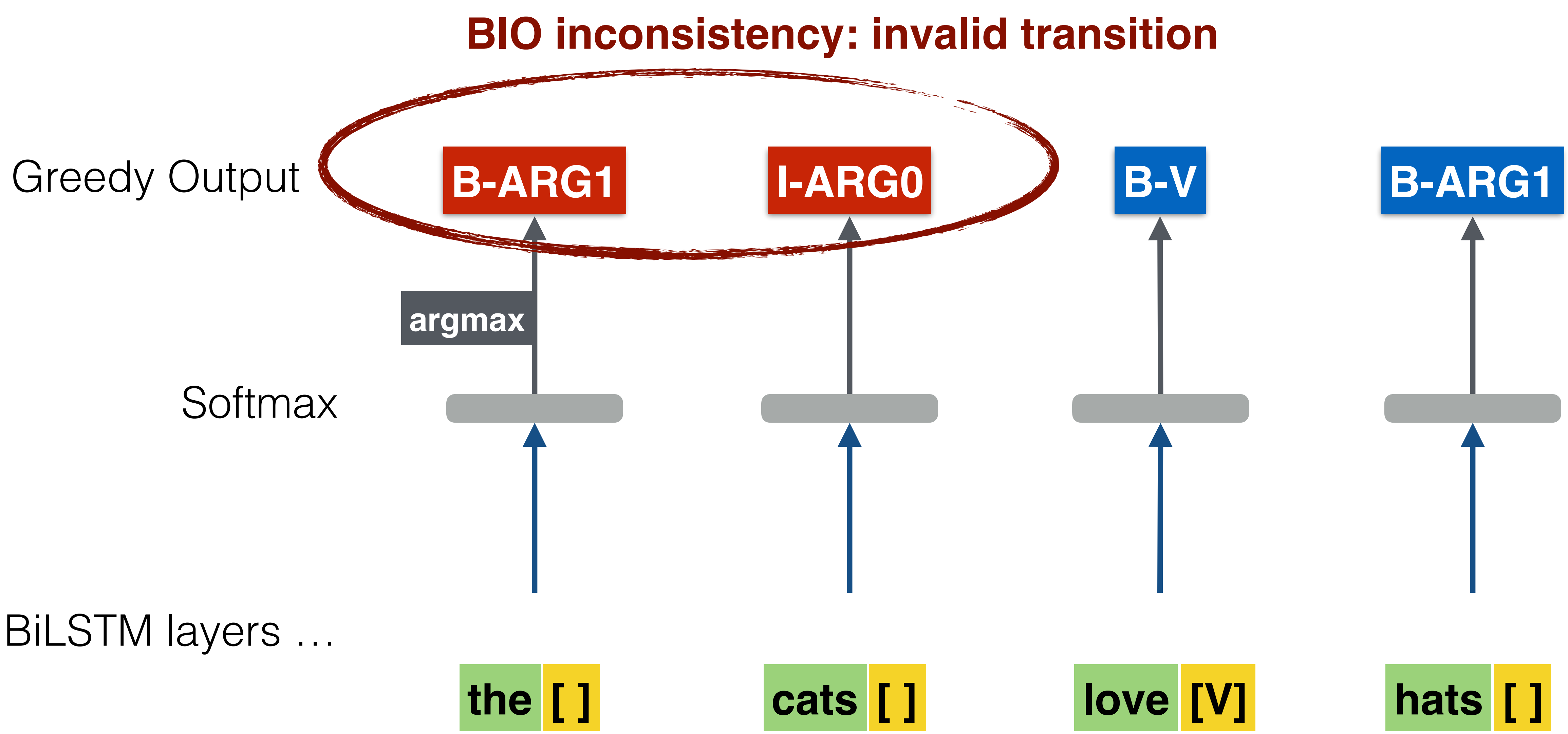
$$\mathbf{r}_t = \sigma(f(\mathbf{h}_{t-1}, \mathbf{x}_t))$$

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

Model - Viterbi Decoding with Hard Constraints



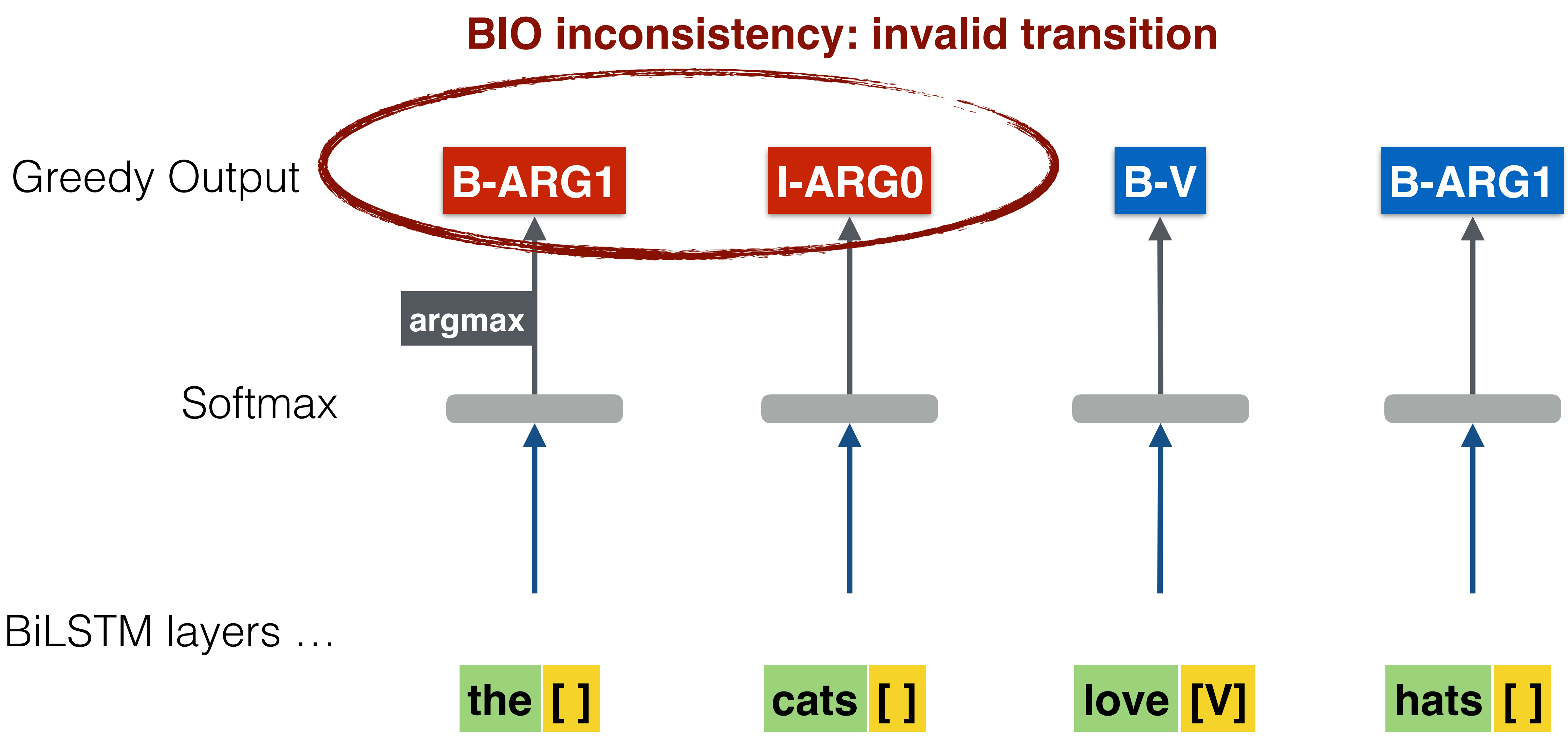
Model - Viterbi Decoding with Hard Constraints



Model - Viterbi Decoding with Hard Constraints

Heuristic transition scores

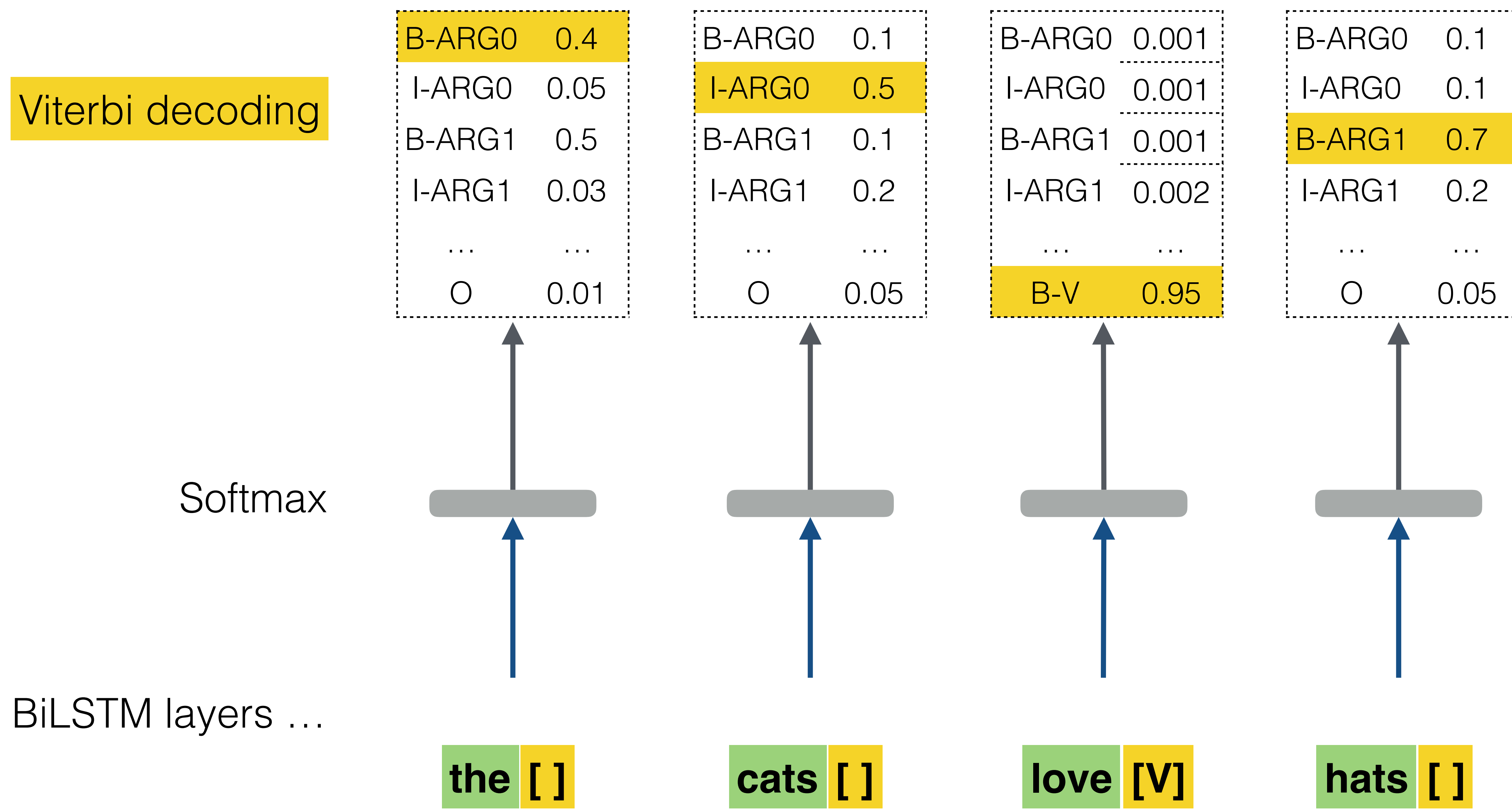
$$s(\text{B-ARG0} \rightarrow \text{I-ARG0}) = 0$$
$$s(\text{B-ARG1} \rightarrow \text{I-ARG0}) = -\inf$$
$$\ldots$$



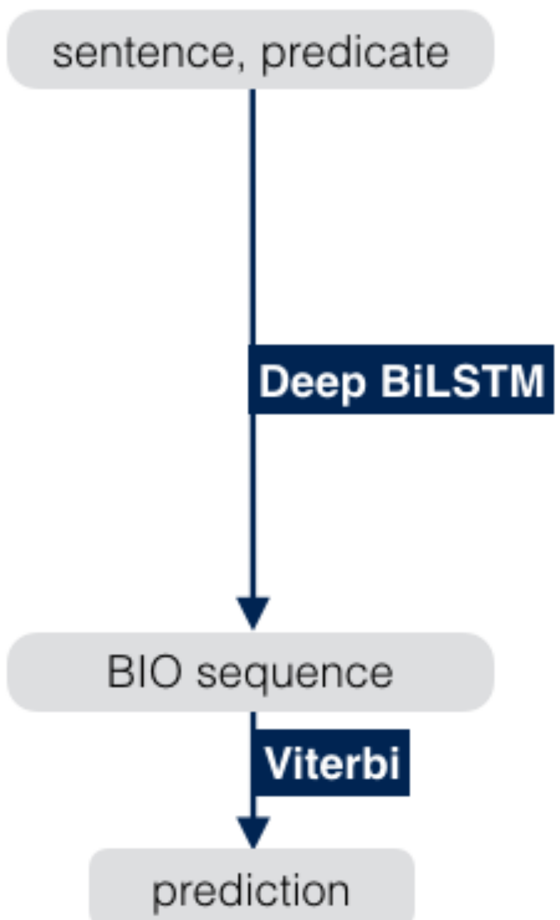
Model - Viterbi Decoding with Hard Constraints

Heuristic transition scores

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



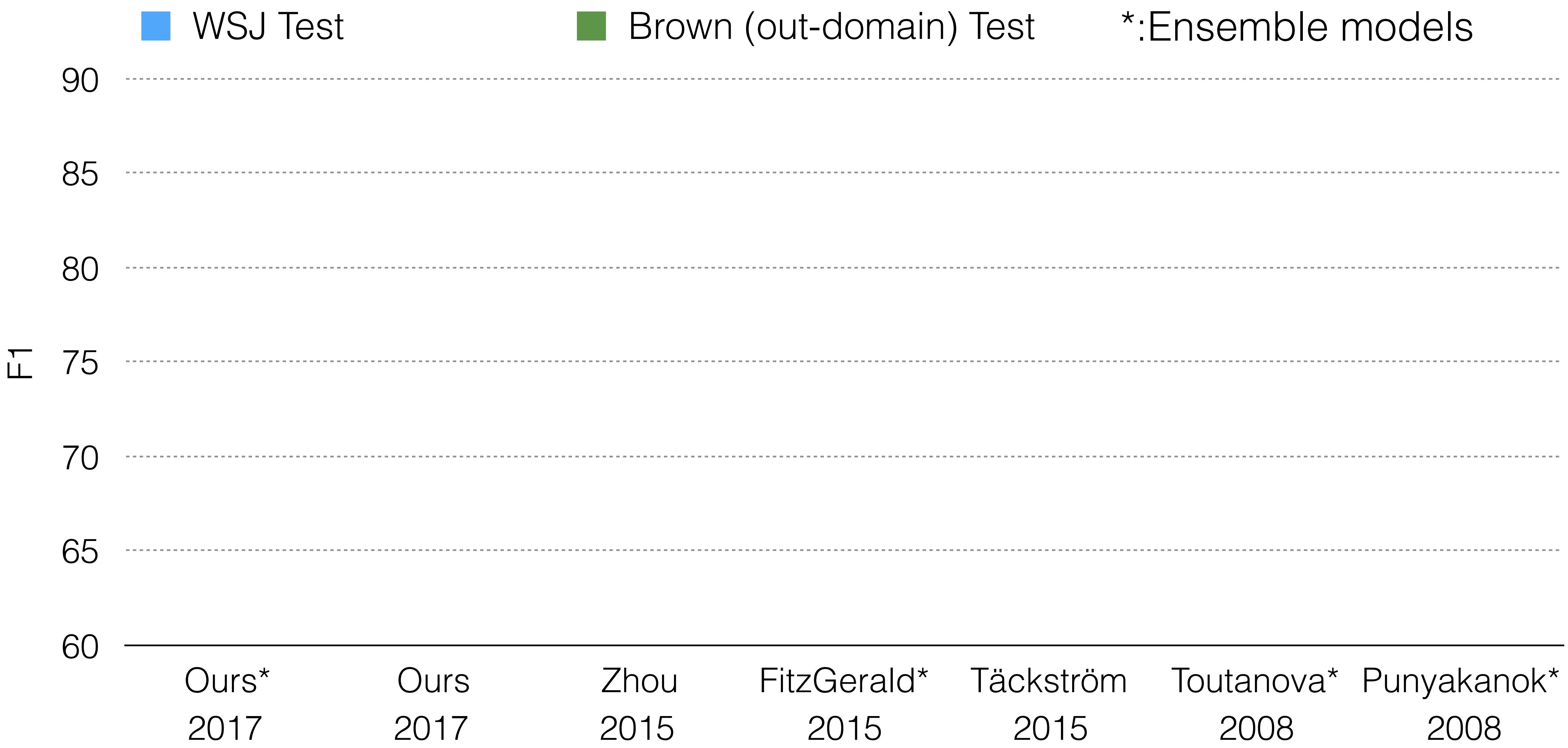
Other Implementation Details …



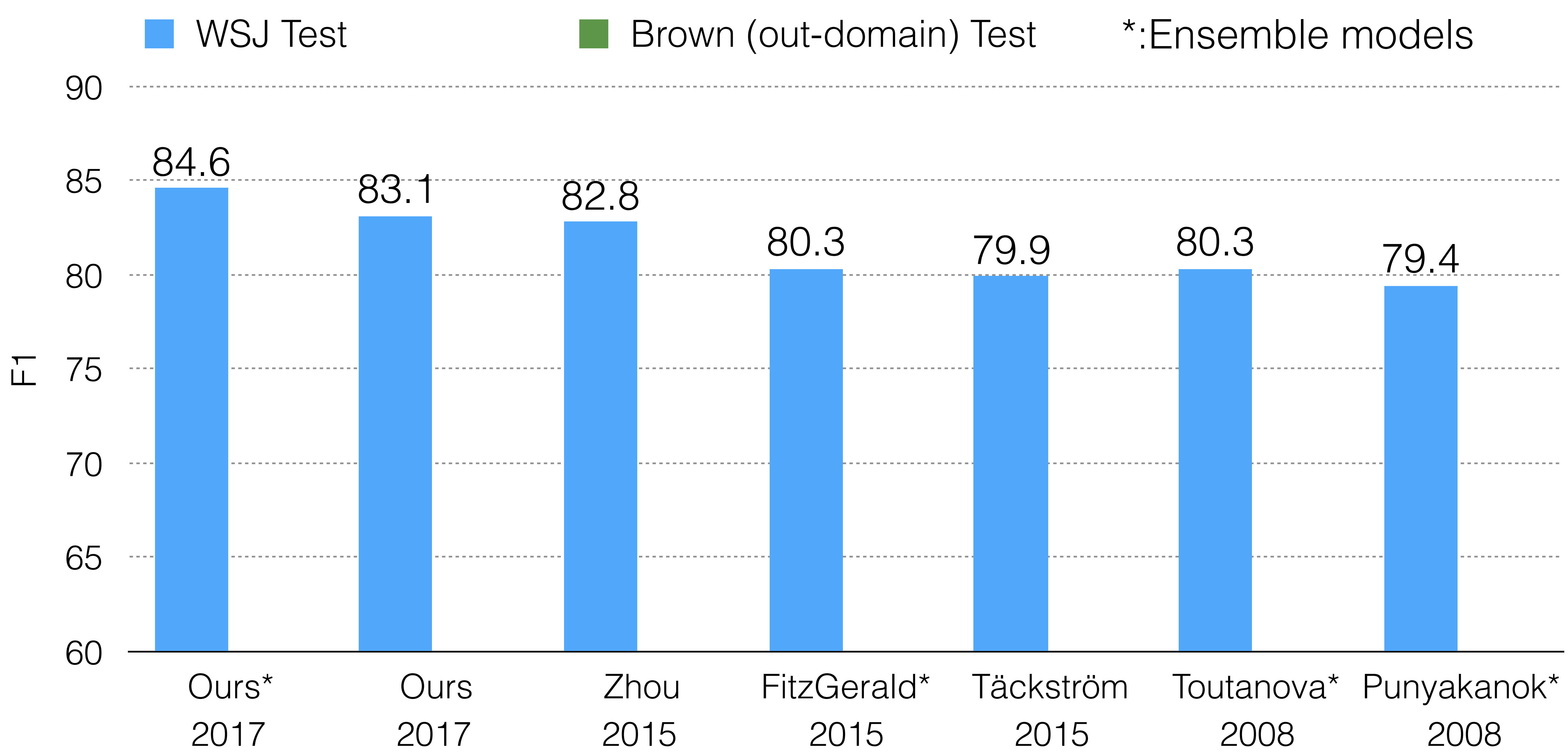
- 8 layer BiLSTMs with 300D hidden layers.
- 100D GloVe embeddings, updated during training.
- **Orthonormal initialization** for LSTM weight matrices (Saxe et al., 2013)
- 0.1 **variational dropout** between layers (Gal and Ghahramani, 2016)
- Trained for 500 epochs.

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

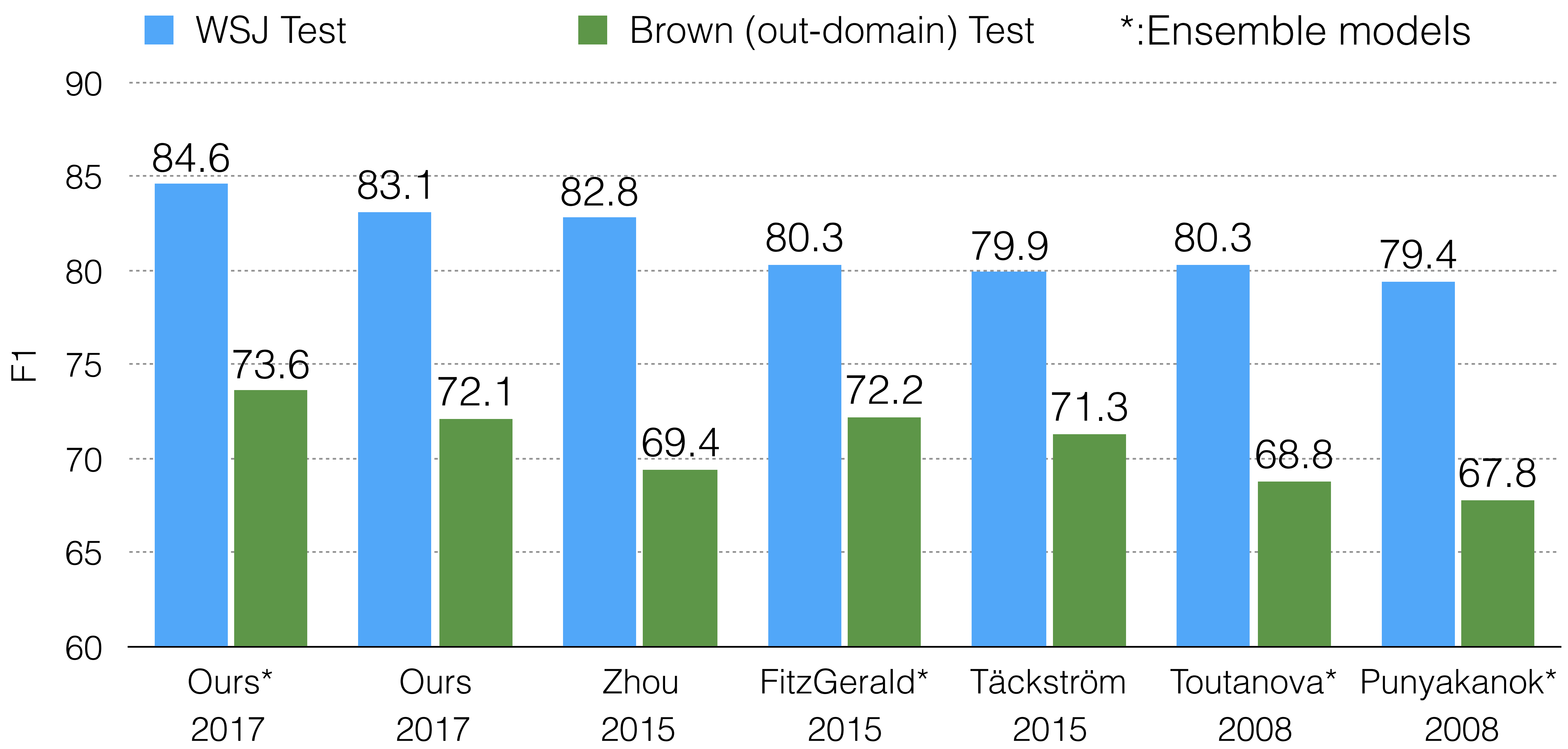
CoNLL 2005 Results



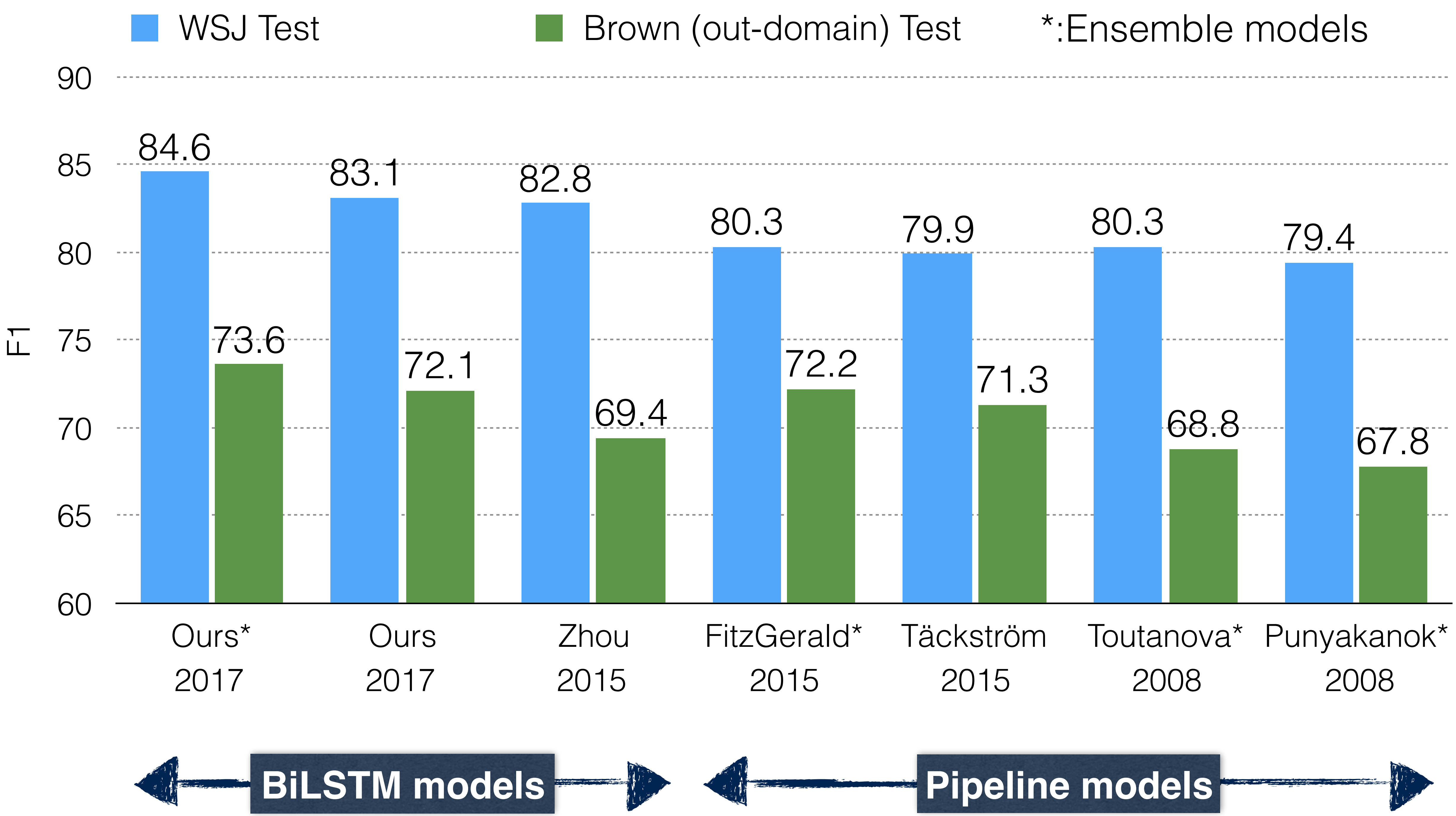
CoNLL 2005 Results



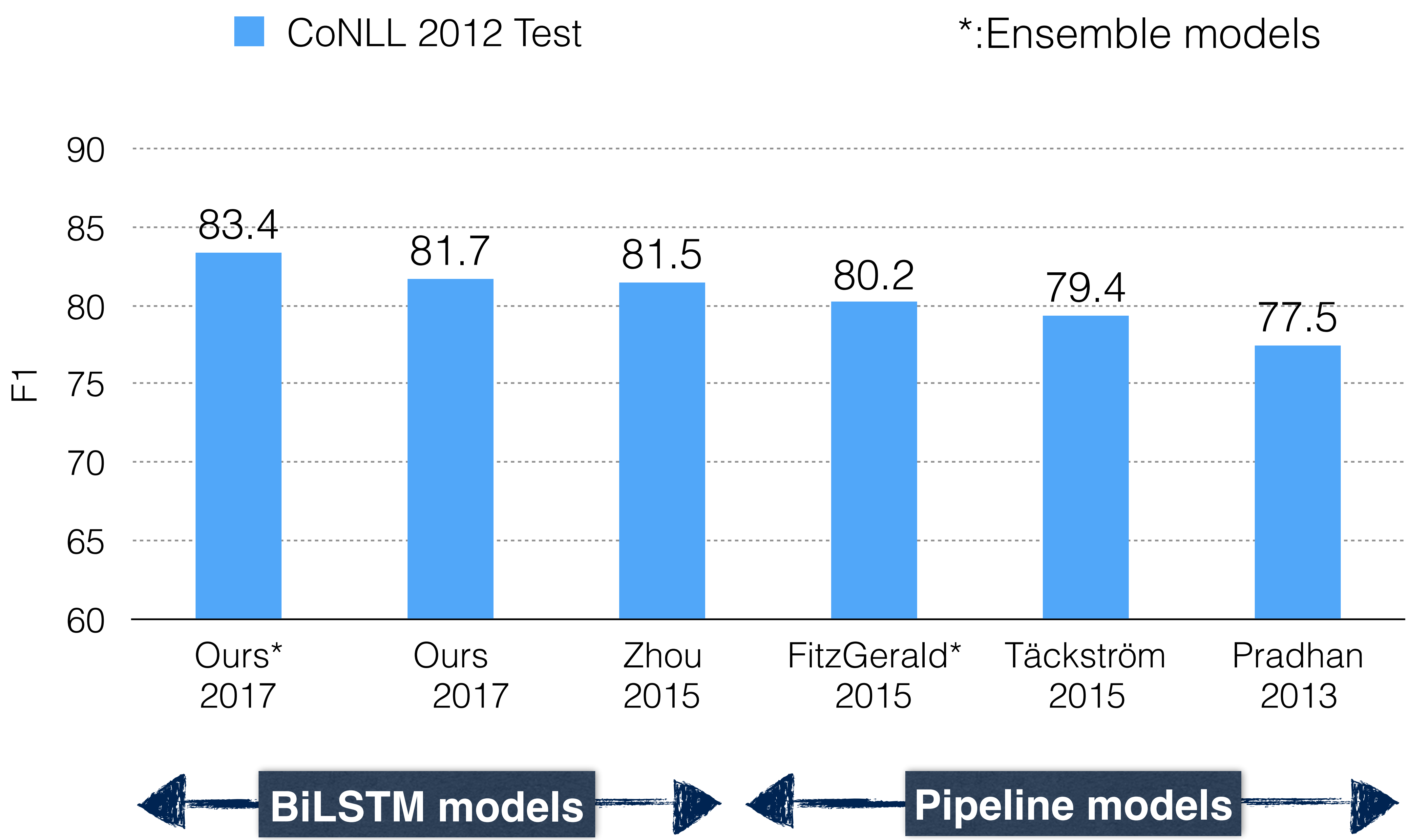
CoNLL 2005 Results



CoNLL 2005 Results

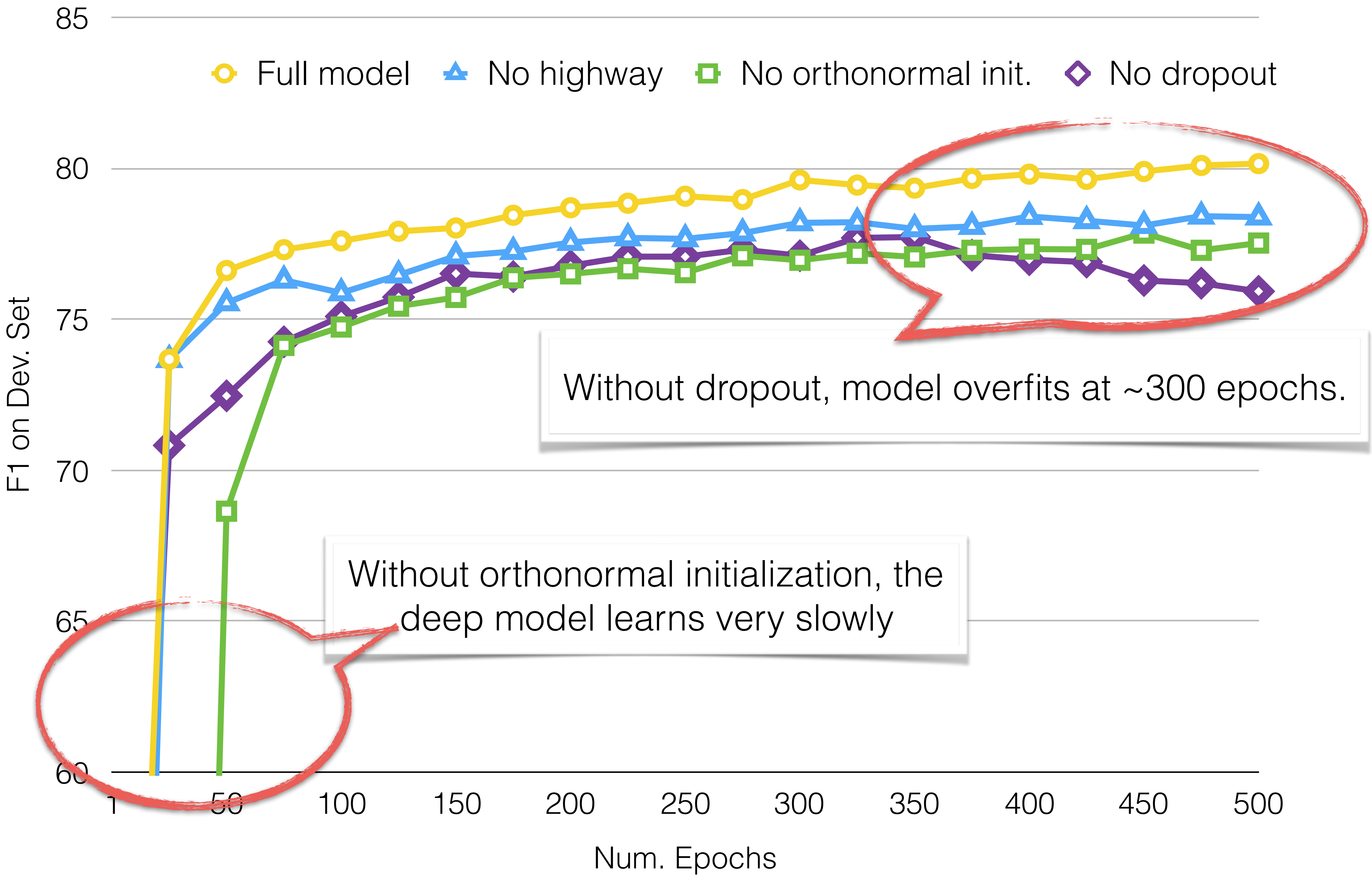


CoNLL 2012 (OntoNotes) Results



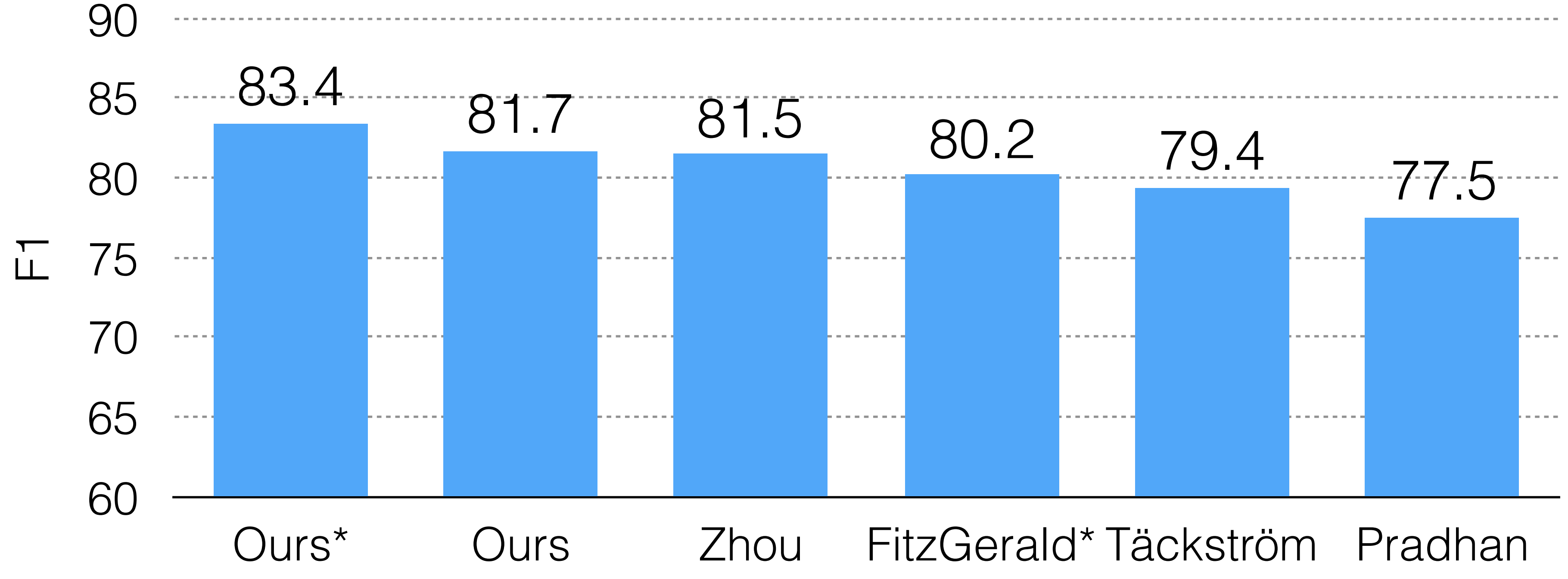
Ablations

(single model, on CoNLL05 Dev)



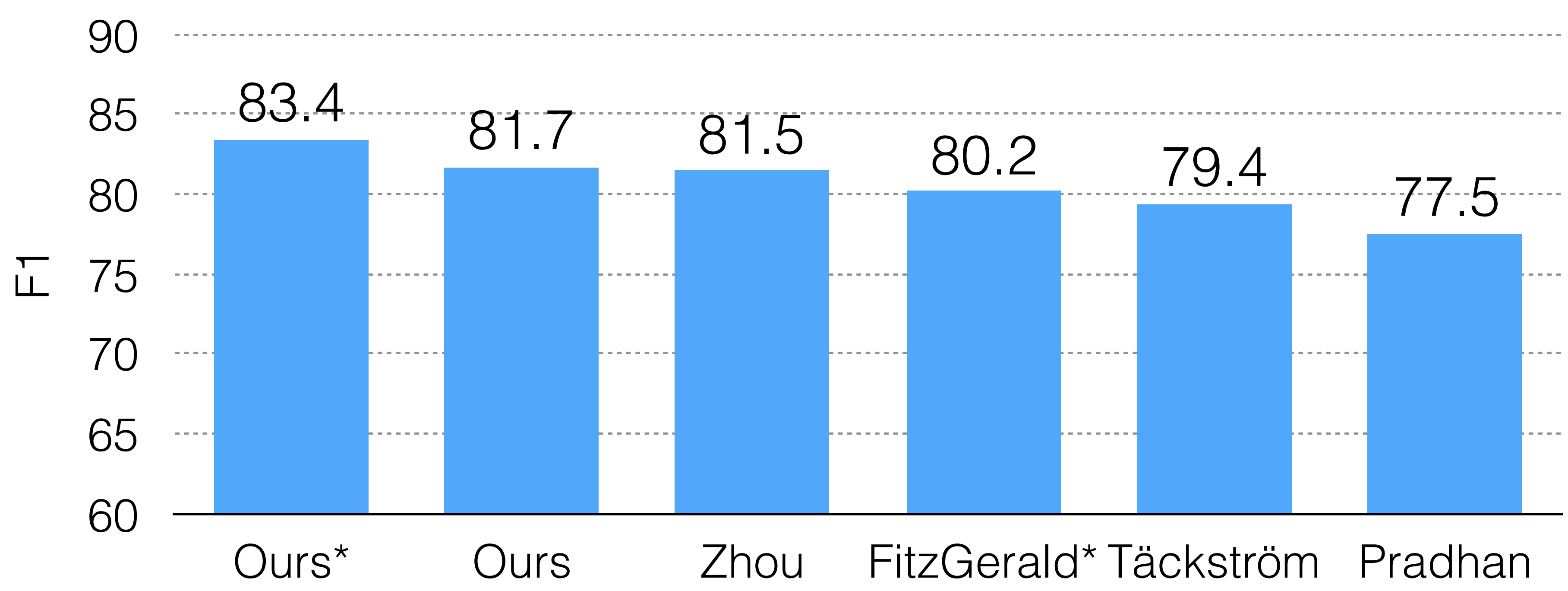
What can we learn from the results?

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. 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?

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. prepositional phrase (PP) attachment

Error Breakdown

Oracle Transformations

Fix Label:

[We] _fly_ to NYC tomorrow.

~~ARG0~~

ARG1

Error Breakdown

Oracle Transformations

Fix Label:

[We] *fly* to NYC tomorrow.

~~ARG0~~

ARG1

Labeling error 29%

Error Breakdown

Oracle Transformations

Fix Label:

[We] *fly* to NYC tomorrow.

~~ARG0~~

ARG1

Labeling error
29%

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

Fix Label:

[We] *fly* to NYC tomorrow.

~~ARG0~~

ARG1

Labeling error
29%

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]

Attachment error
25%

Labeling Errors

Confusion matrix for labeling errors (column normalized)

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	-	55	11	13	4	0	0	0	0	0
A1	78	-	46	0	0	22	11	10	25	14
A2	11	23	-	48	15	56	33	41	25	0
A3	3	2	2	-	4	0	0	0	25	14
ADV	0	0	0	4	-	0	15	29	25	36
DIR	0	0	5	4	0	-	11	2	0	0
LOC	5	9	12	0	4	0	-	10	0	14
MNR	3	0	12	26	33	0	0	-	0	21
PNC	0	3	5	4	0	11	4	2	-	0
TMP	0	8	5	0	41	11	26	6	0	-

Labeling Errors

Confusion matrix for labeling errors (column normalized)

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	-	55	11	13	4	0	0	0	0	0
A1	78	-	46	0	0	22	11	10	25	14
A2	11	23	-	48	15	56	33	41	25	0
A3	3	2	2	-	4	0	0	0	25	14
ADV	0	0	0	4	-	0	15	29	25	36
DIR	0	0	5	4	0	-	11	2	0	0
LOC	5	9	12	0	4	0	-	10	0	14
MNR	3	0	12	26	33	0	0	-	0	21
PNC	0	3	5	4	0	11	4	2	-	0
TMP	0	8	5	0	41	11	26	6	0	-

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

Labeling Errors

Confusion matrix for labeling errors (column normalized)

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	-	55	11	13	4	0	0	0	0	0
A1	78	-	46	0	0	22	11	10	25	14
A2	11	23	-	48	15	56	33	41	25	0
A3	3	2	2	-	4	0	0	0	25	14
ADV	0	0	0	4	-	0	15	29	25	36
DIR	0	0	5	4	0	-	11	2	0	0
LOC	5	9	12	0	4	0	-	10	0	14
MNR	3	0	12	26	33	0	0	-	0	21
PNC	0	3	5	4	0	11	4	2	-	0
TMP	0	8	5	0	41	11	26	6	0	-

- 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 (column normalized)

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	-	55	11	13	4	0	0	0	0	0
A1	78	-	46	0	0	22	11	10	25	14
A2	11	23	-	48	15	56	33	41	25	0
A3	3	2	2	-	4	0	0	0	25	14
ADV	0	0	0	4	-	0	15	29	25	36
DIR	0	0	5	4	0	-	11	2	0	0
LOC	5	9	12	0	4	0	-	10	0	14
MNR	3	0	12	26	33	0	0	-	0	21
PNC	0	3	5	4	0	11	4	2	-	0
TMP	0	8	5	0	41	11	26	6	0	-

- 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 expert annotators!

PP Attachment

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

PP Attachment

Wrong PP attachment
(attach high)

Arg1 (NP) Arg2 (PP)

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

Arg1 (NP)

Correct PP attachment
(attach low)

PP Attachment

Wrong PP attachment
(attach high)

Arg1 (NP) Arg2 (PP)

Wrong SRL spans

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

merge

Arg1 (NP)

Correct SRL spans

Correct PP attachment
(attach low)

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

Attachment mistakes: 25%.

Categorize the Y spans in :
[XY]—>[X]**[Y]** and
[X]**[Y]**—>[XY] operations
by 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)

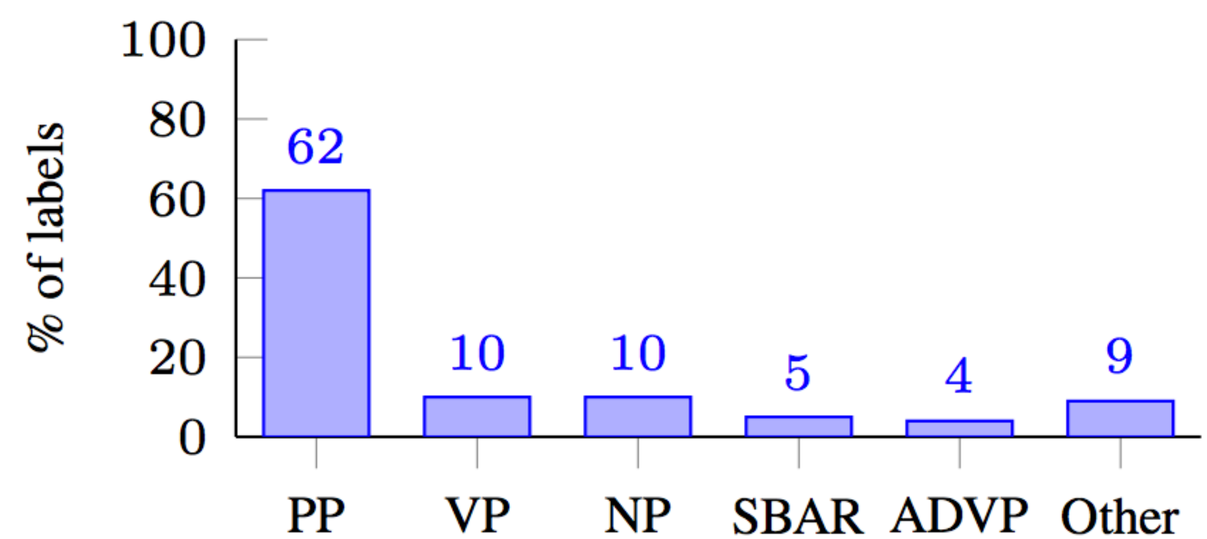
Wrong SRL spans

merge

Correct SRL spans

Attachment mistakes: 25%.

Categorize the Y spans in :
[XY]—>[X]**[Y]** and
[X]**[Y]**—>[XY] operations
by 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

Takeaway

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

— Use external information/PropBank frame inventory.

Question (2): 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 (unlabeled).

— See if SRL performance increases.

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

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

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] $\in$ Syntax Tree

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

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

ARG0 ARG1

Penalize sequence score

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

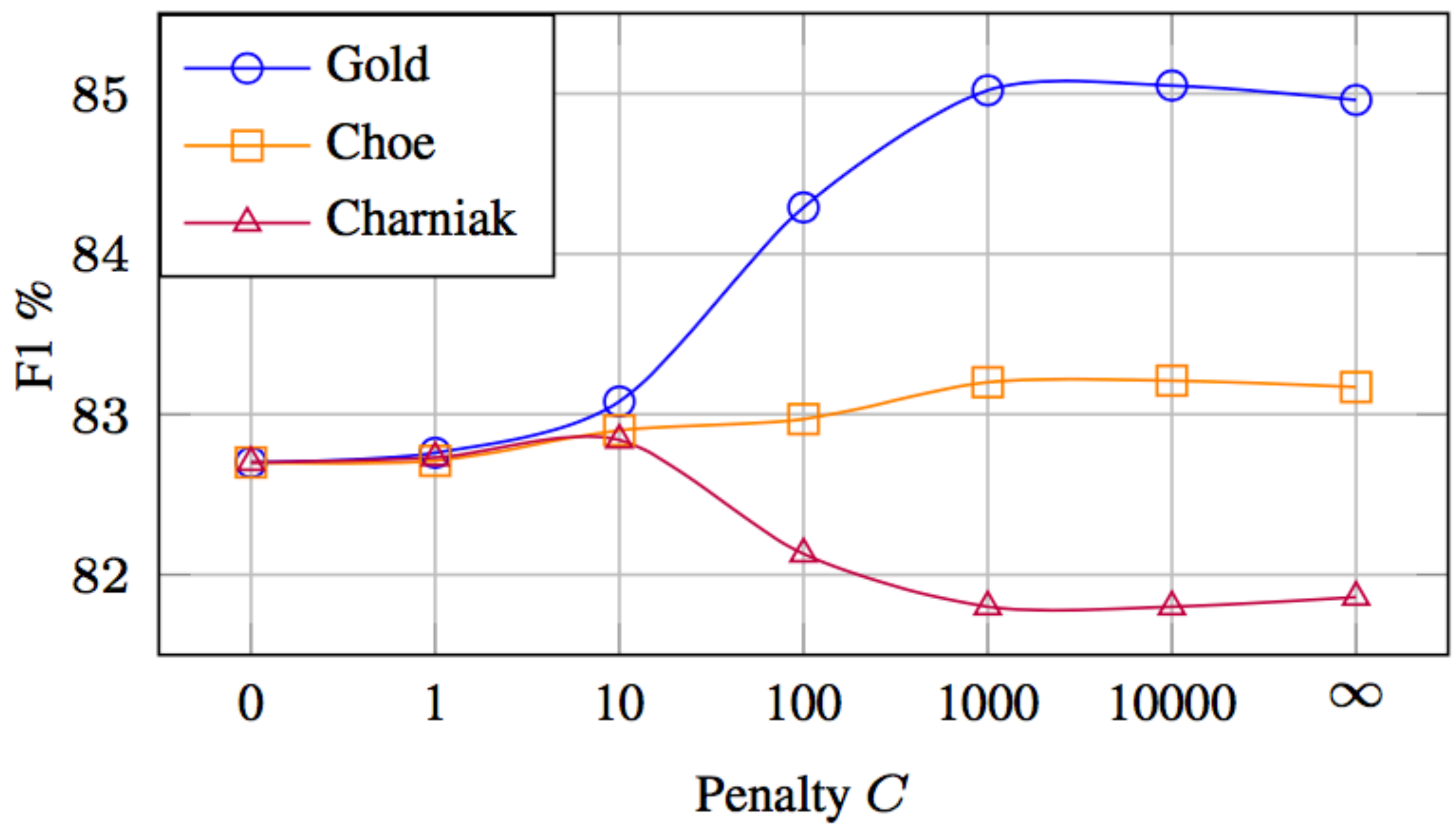
Penalty strength

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?

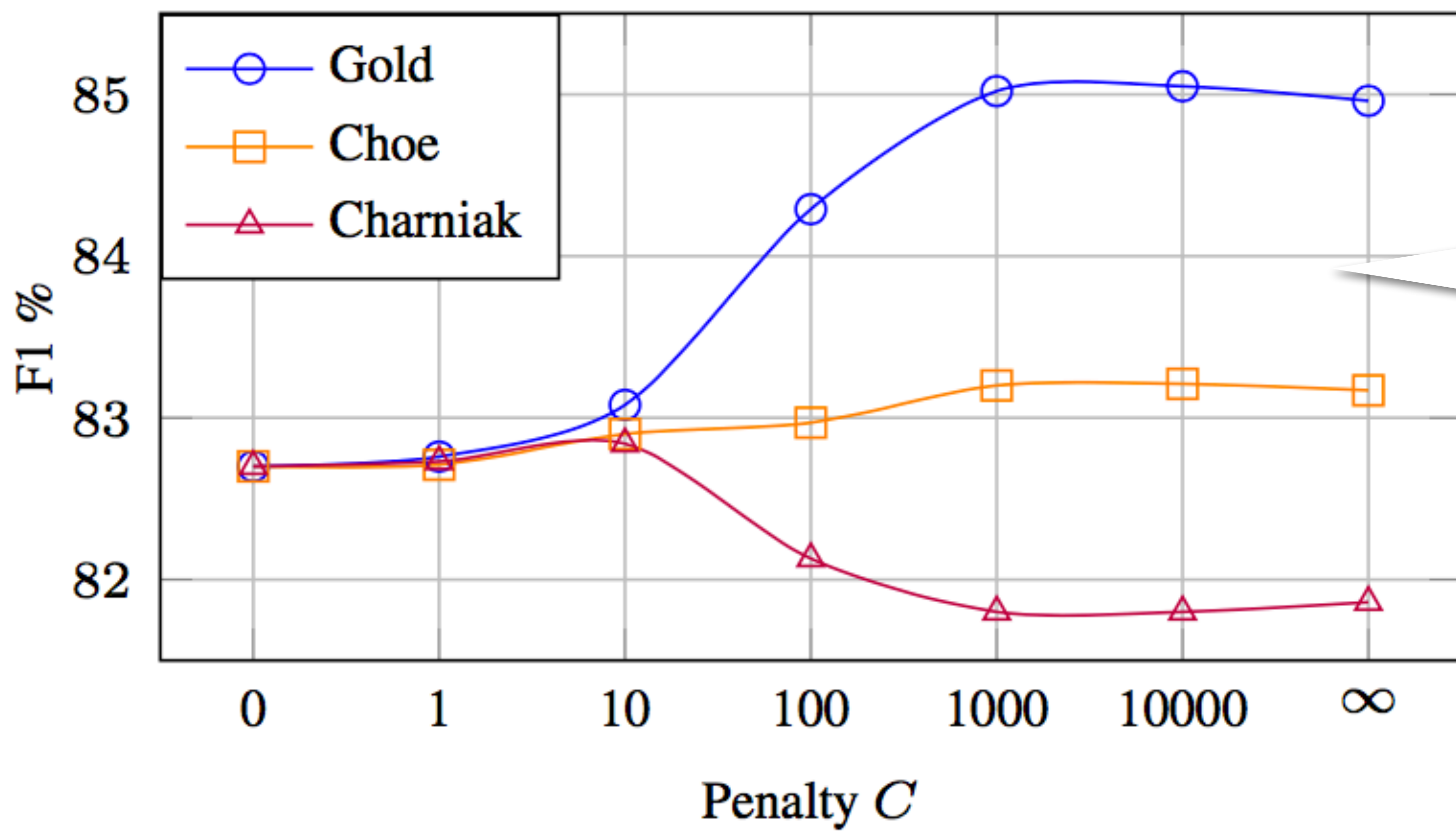
Syntax Decoding Results



Gold: Penn Treebank constituents.
Choe: Parsing as language modeling, Choe and Charniak, 2016 (SOTA)
Charniak: A maximum-entropy-inspired parser, Charniak, 2000

Can Syntax Still Help?

Syntax Decoding Results



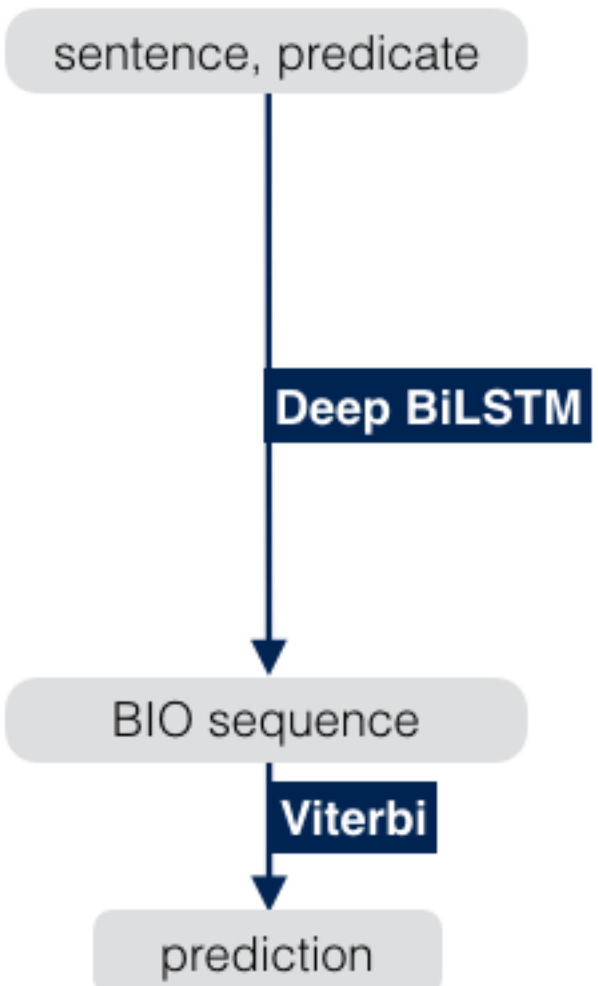
Takeaway

— Modest gain when using accurate syntax.

— More improvement: Joint training, use syntactic labels, etc.

Gold: Penn Treebank constituents.
Choe: Parsing as language modeling, Choe and Charniak, 2016 (SOTA)
Charniak: A maximum-entropy-inspired parser, Charniak, 2000

Thank You!



- New state-of-the-art deep network for end-to-end SRL.
- Code and models are 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.