Graph-based Methods

Naive Baseline: Splitting into chunks

```
In February
1885 Gordon
returned to the
Sudan to
evacuate
Egyptian
forces.
Khartoum came
under siege the
next month and
rebels broke
into the city,
...
```

Naive Baseline: Splitting into chunks

Split

```
In February
1885 Gordon
returned to the
Sudan to
evacuate
Egyptian
forces.
Khartoum came
under siege the
next month and
rebels broke
into the city,
...
```

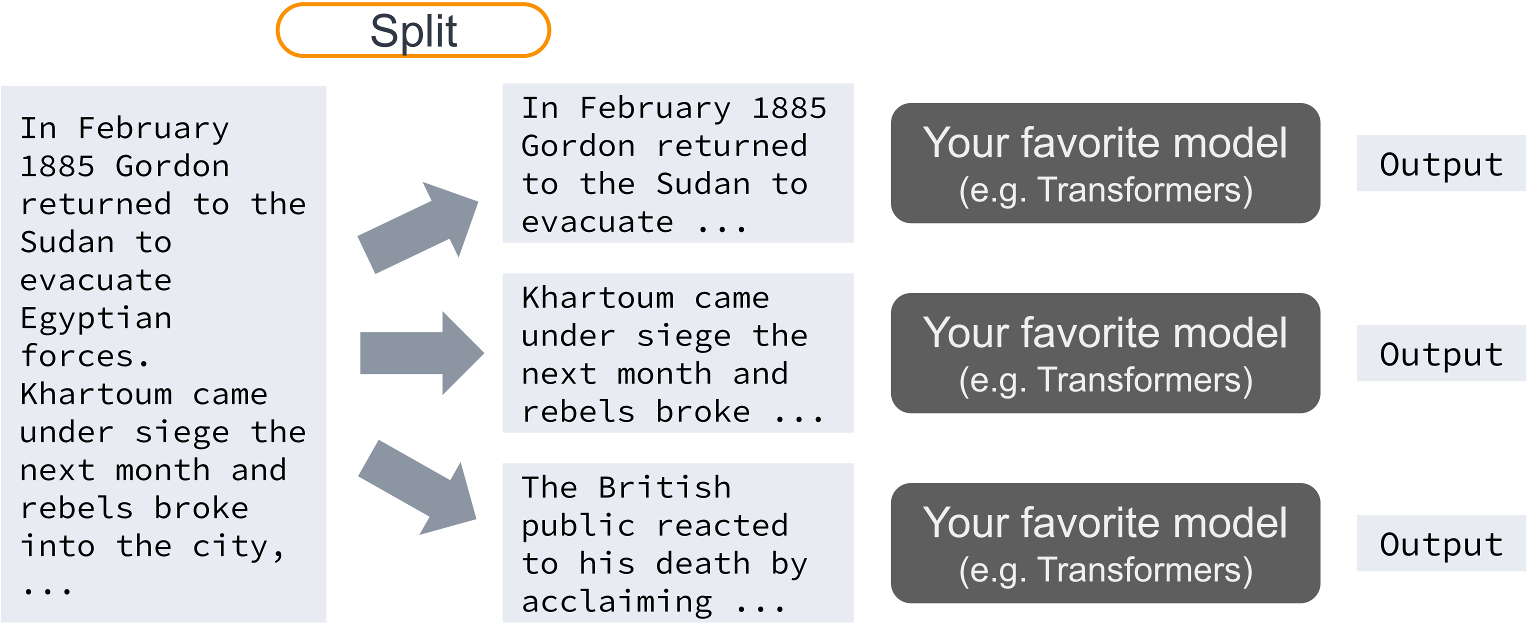
```
In February 1885
Gordon returned
to the Sudan to
evacuate ...
```

```
Khartoum came
under siege the
next month and
rebels broke ...
```

```
The British
public reacted
to his death by
acclaiming ...
```

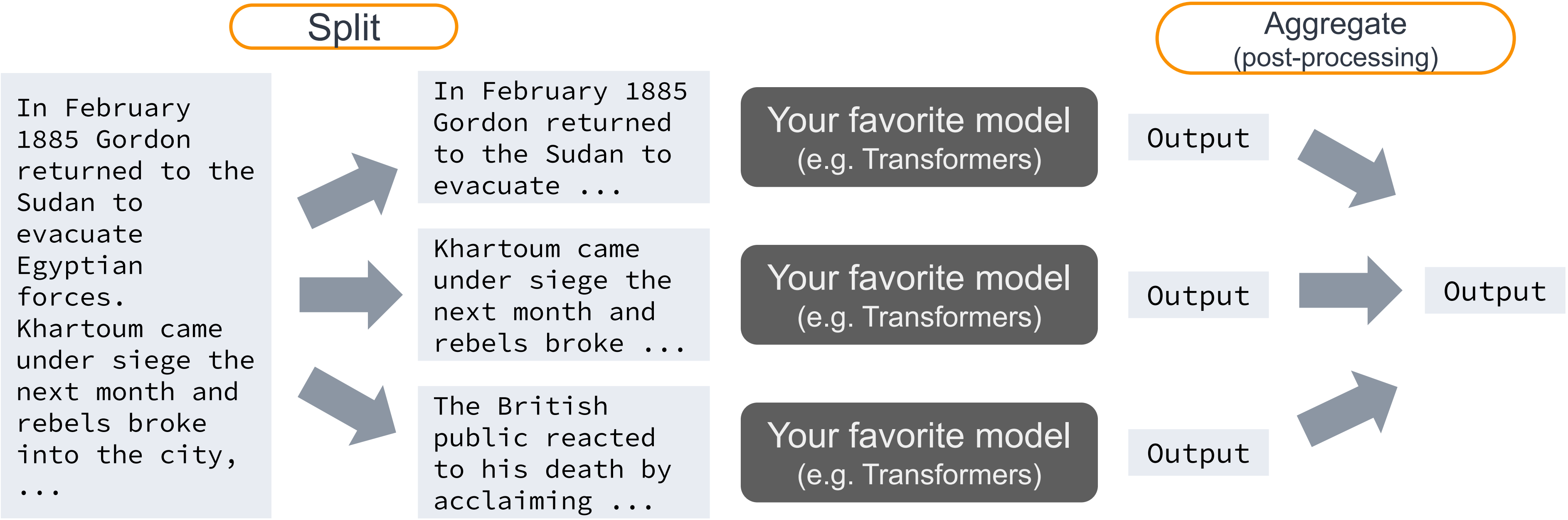
Clark & Gardner, "Simple and Effective Multi-Paragraph Reading Comprehension." 2017.

Naive Baseline: Splitting into chunks



Clark & Gardner, “Simple and Effective Multi-Paragraph Reading Comprehension.” 2017.

Naive Baseline: Splitting into chunks



Clark & Gardner, “Simple and Effective Multi-Paragraph Reading Comprehension.” 2017.

Naive Baseline: Splitting into chunks

Split

Aggregate (post-processing)

```
In February 1885 Gordon returned to the Sudan to evacuate Egyptian forces. Khartoum came under siege the next month and rebels broke into the city, ...
```

```
In February 1885 Gordon returned to the Sudan to evacuate ...
```

Your favorite model (e.g. Transformers) → `Output`

```
Khartoum came under siege the next month and rebels broke ...
```

Your favorite model (e.g. Transformers) → `Output`

```
The British public reacted to his death by acclaiming ...
```

Your favorite model (e.g. Transformers) → `Output`

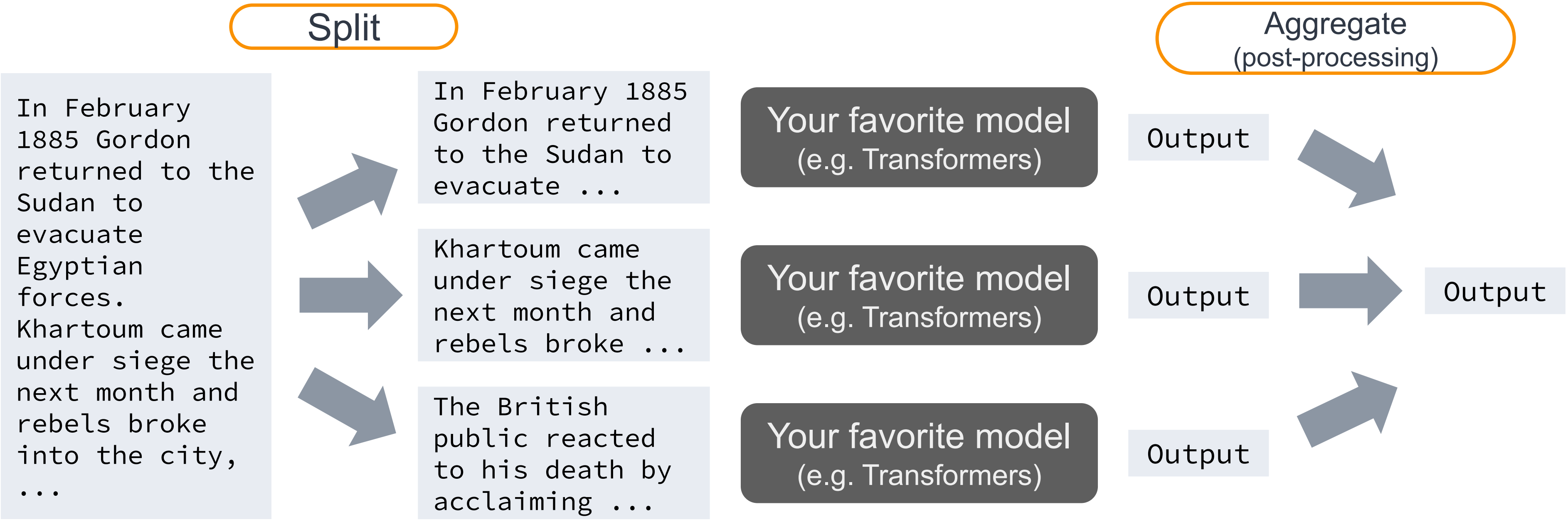
`Output`

Strong baselines - See Open-domain QA Tutorial!

Clark & Gardner, "Simple and Effective Multi-Paragraph Reading Comprehension." 2017.

UWNLP AI2

Naive Baseline: Splitting into chunks



Limited when long-range dependencies are needed

Clark & Gardner, “Simple and Effective Multi-Paragraph Reading Comprehension.” 2017.

Overview

- Hierarchical modeling
- Graph-based modeling
- Graph-based modeling w/ external knowledge

Hierarchical Modeling

- Leverage natural hierarchy of the document (words→sentences→paragraphs)

```
Pork belly =
delicious.
Scallops? These
were amazing.
Fun and tasty
cocktails. Next
time I in
Phoenix, I will
go back here.
Highly
recommended.
```

Document

Yang et al. “Hierarchical Attention Network for Document Classification”. 2016.

Hierarchical Modeling

- Leverage natural hierarchy of the document (words→sentences→paragraphs)

Document	Sentences
Pork belly = delicious. Scallops? These were amazing. Fun and tasty cocktails. Next time I in Phoenix, I will go back here. Highly recommended.	Pork belly = delicious. These were amazing. Next time I in Phoenix, I will go back here. Highly recommended.

Yang et al. “Hierarchical Attention Network for Document Classification”. 2016.

Hierarchical Modeling

- Leverage natural hierarchy of the document (words→sentences→paragraphs)

Document	Sentences	Words
Pork belly = delicious. Scallops? These were amazing. Fun and tasty cocktails. Next time I in Phoenix, I will go back here. Highly recommended.	Pork belly = delicious.	Pork / belly / = / delicious
	These were amazing.	These / were / amazing
	Next time I in Phoenix, I will go back here.	Next / Time / I / in / Phoenix
	Highly recommended.	Highly / recommended

Yang et al. "Hierarchical Attention Network for Document Classification". 2016.

UWNLP AI2

Hierarchical Modeling

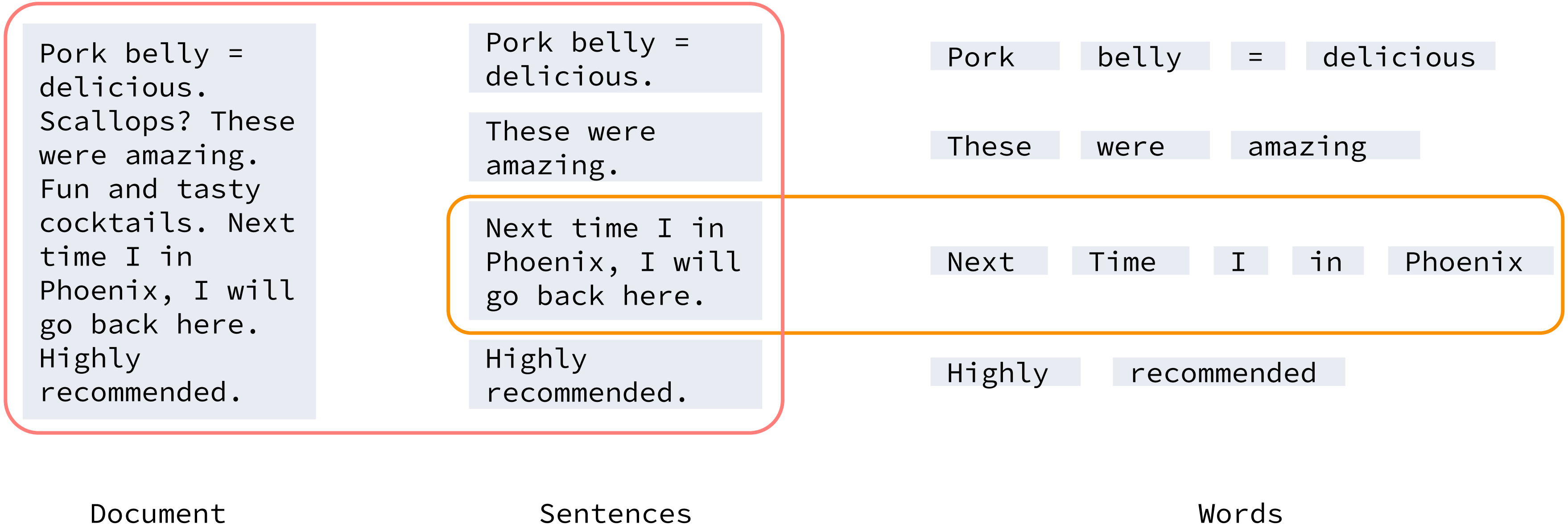
- Leverage natural hierarchy of the document (words→sentences→paragraphs)

Document	Sentences	Words
Pork belly = delicious. Scallops? These were amazing. Fun and tasty cocktails. Next time I in Phoenix, I will go back here. Highly recommended.	Pork belly = delicious.	Pork · belly · = · delicious
	These were amazing.	These · were · amazing
	Next time I in Phoenix, I will go back here.	Next · Time · I · in · Phoenix
	Highly recommended.	Highly · recommended

Yang et al. “Hierarchical Attention Network for Document Classification”. 2016.

Hierarchical Modeling

- Leverage natural hierarchy of the document (words→sentences→paragraphs)



Yang et al. “Hierarchical Attention Network for Document Classification”. 2016.

Hierarchical Modeling

- Leverage natural hierarchy of the document (words→sentences→paragraphs)

```
Pork belly =
delicious.
Scallops? These
were amazing.
Fun and tasty
cocktails. Next
time I in
Phoenix, I will
go back here.
Highly
recommended.
```

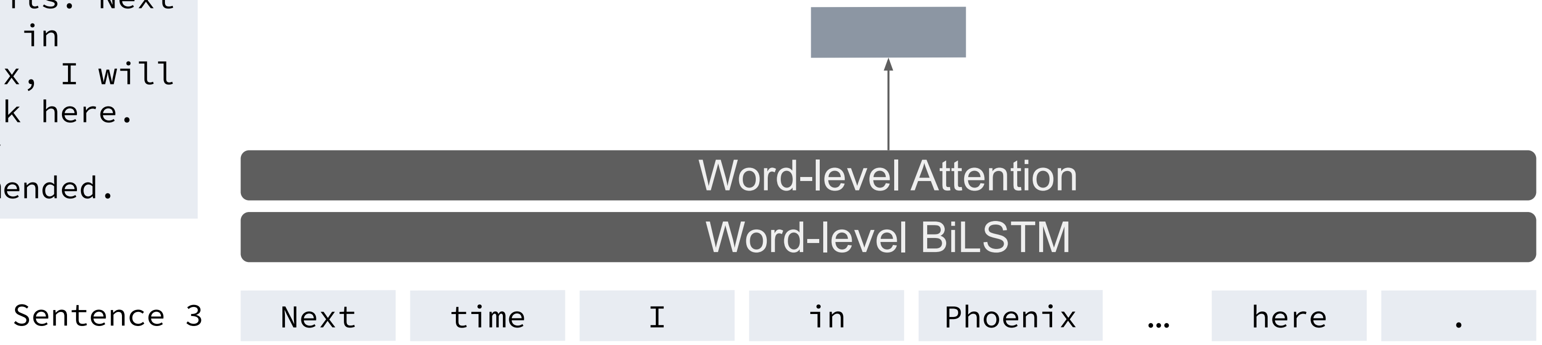
Sentence 3	Next	time	I	in	Phoenix	...	here	.

Yang et al. "Hierarchical Attention Network for Document Classification". 2016.

Hierarchical Modeling

- Leverage natural hierarchy of the document (words→sentences→paragraphs)

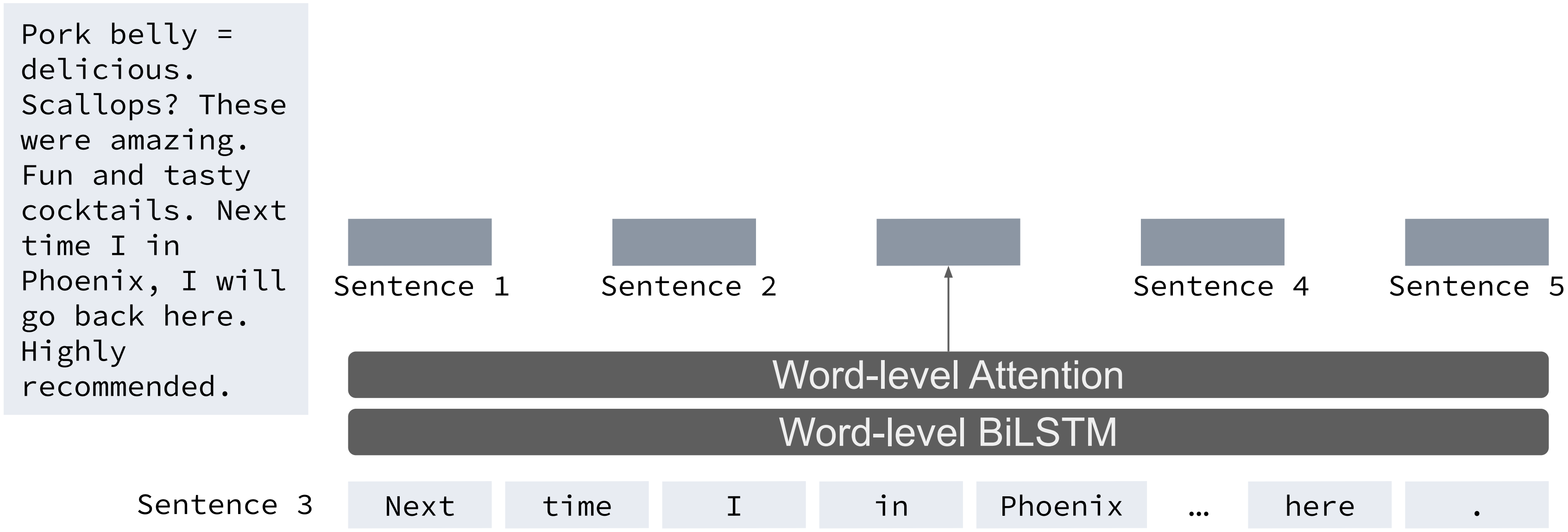
```
Pork belly =
delicious.
Scallops? These
were amazing.
Fun and tasty
cocktails. Next
time I in
Phoenix, I will
go back here.
Highly
recommended.
```



Yang et al. “Hierarchical Attention Network for Document Classification”. 2016.

Hierarchical Modeling

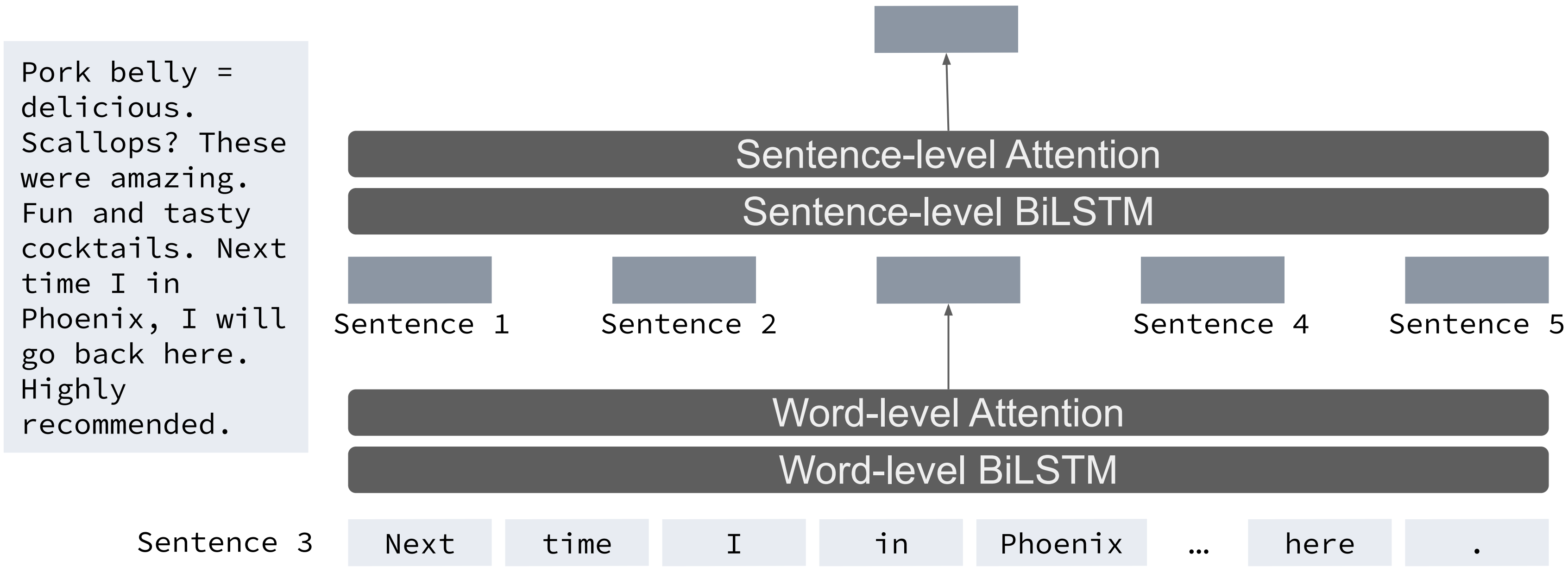
- Leverage natural hierarchy of the document (words→sentences→paragraphs)



Yang et al. "Hierarchical Attention Network for Document Classification". 2016.

Hierarchical Modeling

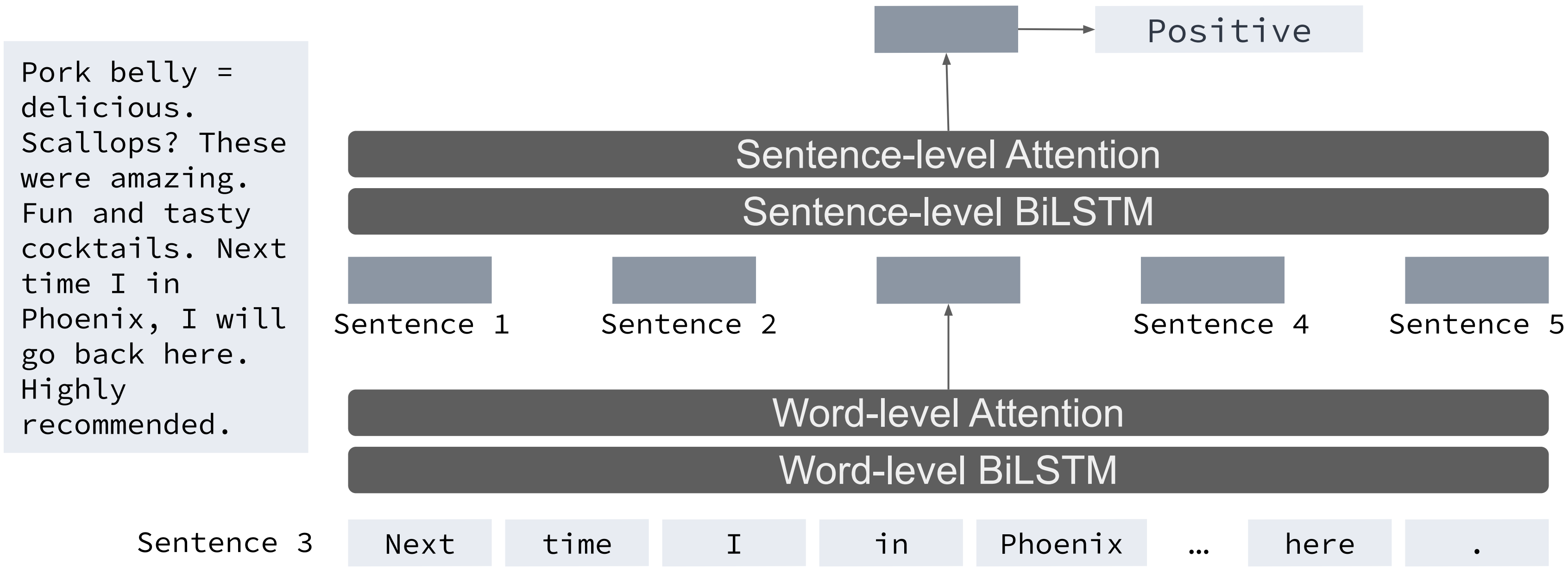
- Leverage natural hierarchy of the document (words→sentences→paragraphs)



Yang et al. “Hierarchical Attention Network for Document Classification”. 2016.

Hierarchical Modeling

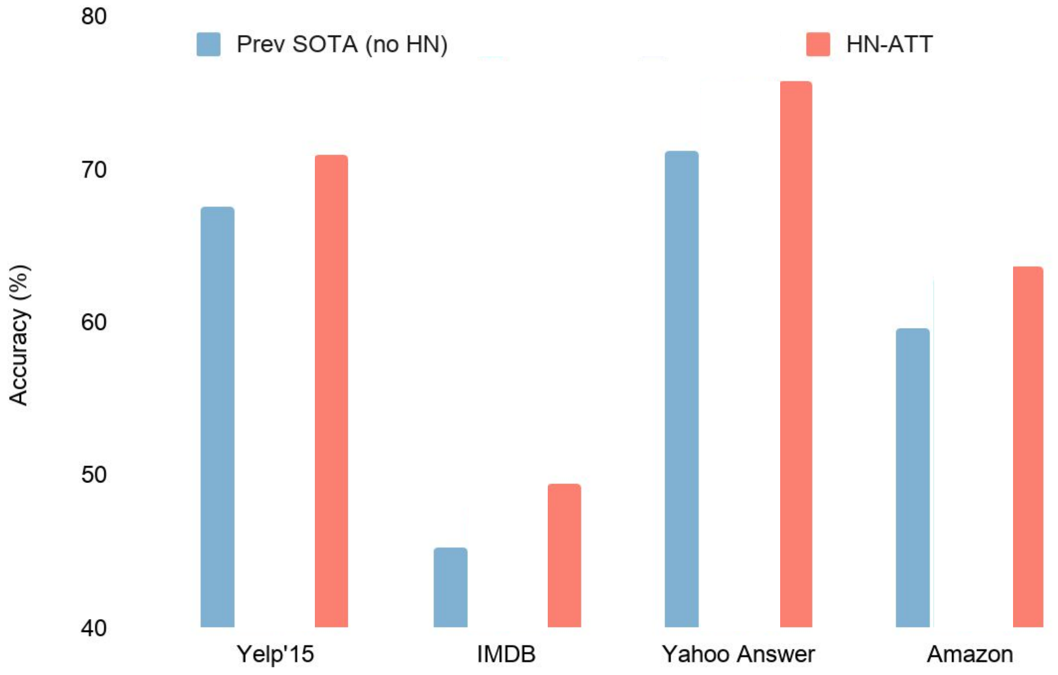
- Leverage natural hierarchy of the document (words→sentences→paragraphs)



Yang et al. "Hierarchical Attention Network for Document Classification". 2016.

Hierarchical Modeling

- Leverage natural hierarchy of the document (words→sentences→paragraphs)



Yang et al. "Hierarchical Attention Network for Document Classification". 2016.

UWNLP AI2

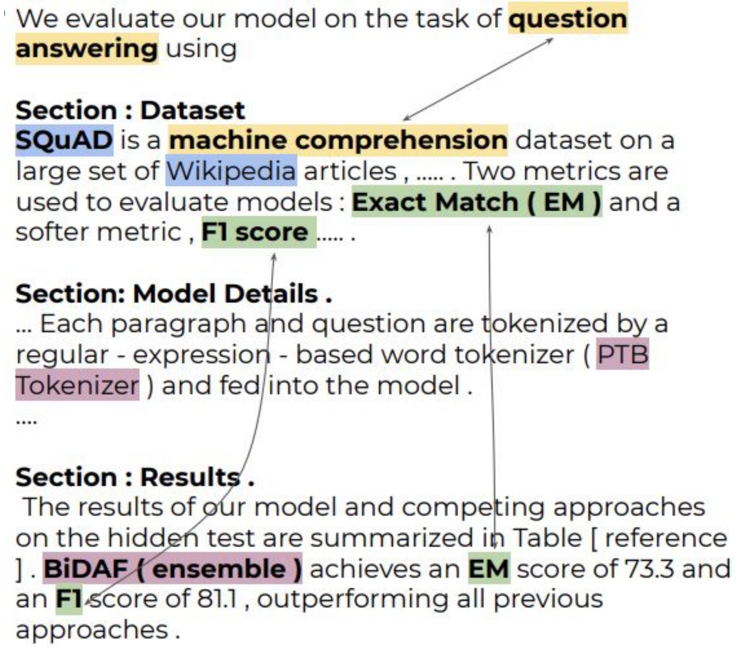
Hierarchical Modeling

- Leverage natural hierarchy of the document (words→sentences→paragraphs)

Yang et al. “Hierarchical Attention Network for Document Classification”. 2016.

Hierarchical Modeling

- Can be used jointly with pretrained Transformers (paragraph→document)

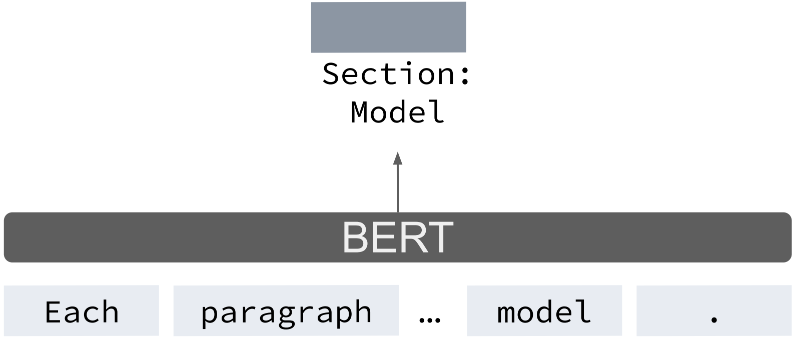


We evaluate our model on the task of **question answering** using

Section : Dataset
SQuAD is a **machine comprehension** dataset on a large set of Wikipedia articles , ... Two metrics are used to evaluate models : **Exact Match (EM)** and a softer metric , **F1 score** ..... .

Section: Model Details .
... Each paragraph and question are tokenized by a regular - expression - based word tokenizer (PTB Tokenizer) and fed into the model .
....

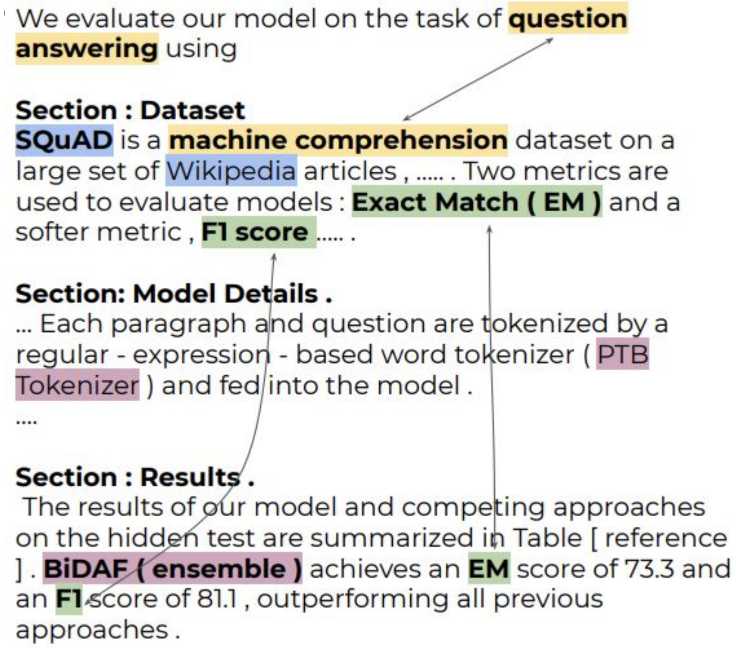
Section : Results .
The results of our model and competing approaches on the hidden test are summarized in Table [reference] . **BiDAF (ensemble)** achieves an **EM** score of 73.3 and an **F1** score of 81.1 , outperforming all previous approaches .



Jain et al. "SciREX: A challenging dataset for Document-level Information Extraction." 2020.

Hierarchical Modeling

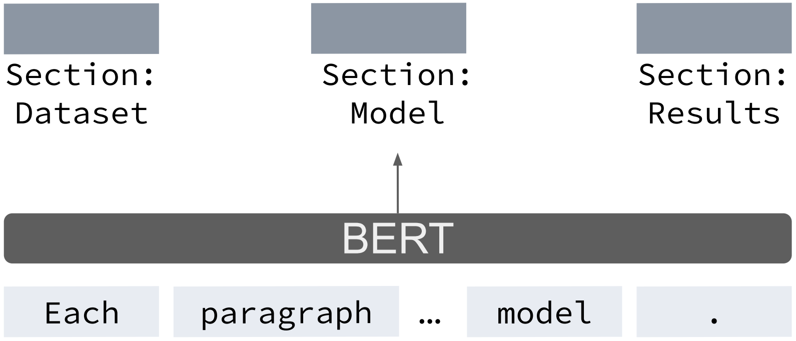
- Can be used jointly with pretrained Transformers (paragraph→document)


We evaluate our model on the task of **question answering** using

Section : Dataset
SQuAD is a **machine comprehension** dataset on a large set of Wikipedia articles , ... Two metrics are used to evaluate models : **Exact Match (EM)** and a softer metric , **F1 score** ..... .

Section: Model Details .
... Each paragraph and question are tokenized by a regular - expression - based word tokenizer (PTB Tokenizer) and fed into the model .
....

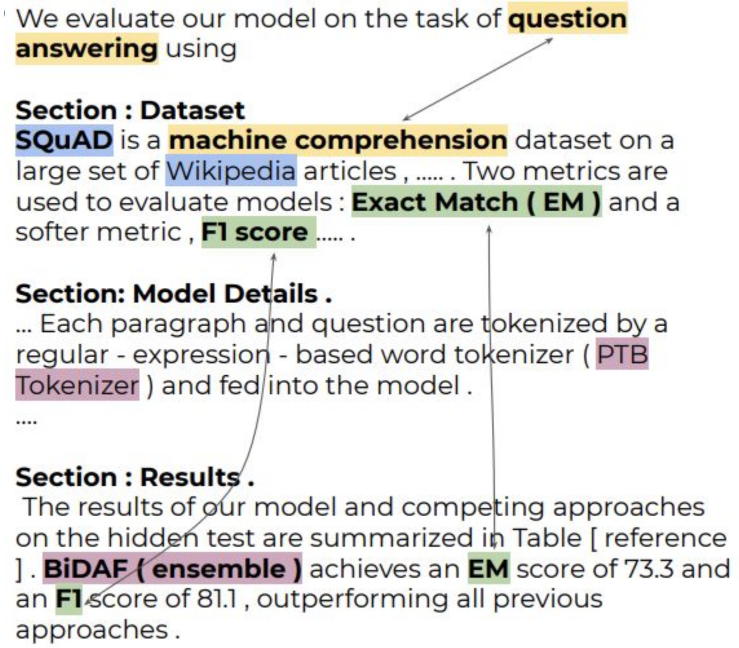
Section : Results .
The results of our model and competing approaches on the hidden test are summarized in Table [reference] . **BiDAF (ensemble)** achieves an **EM** score of 73.3 and an **F1** score of 81.1 , outperforming all previous approaches .

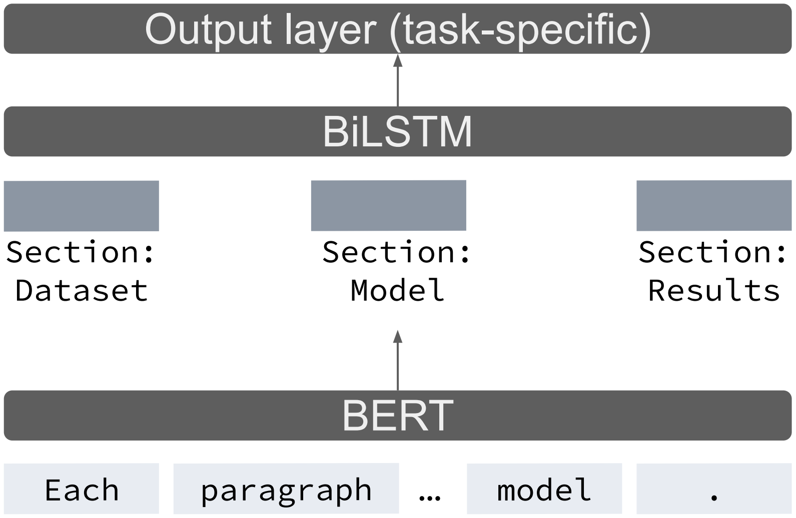


Jain et al. "SciREX: A challenging dataset for Document-level Information Extraction." 2020.

Hierarchical Modeling

- Can be used jointly with pretrained Transformers (paragraph→document)





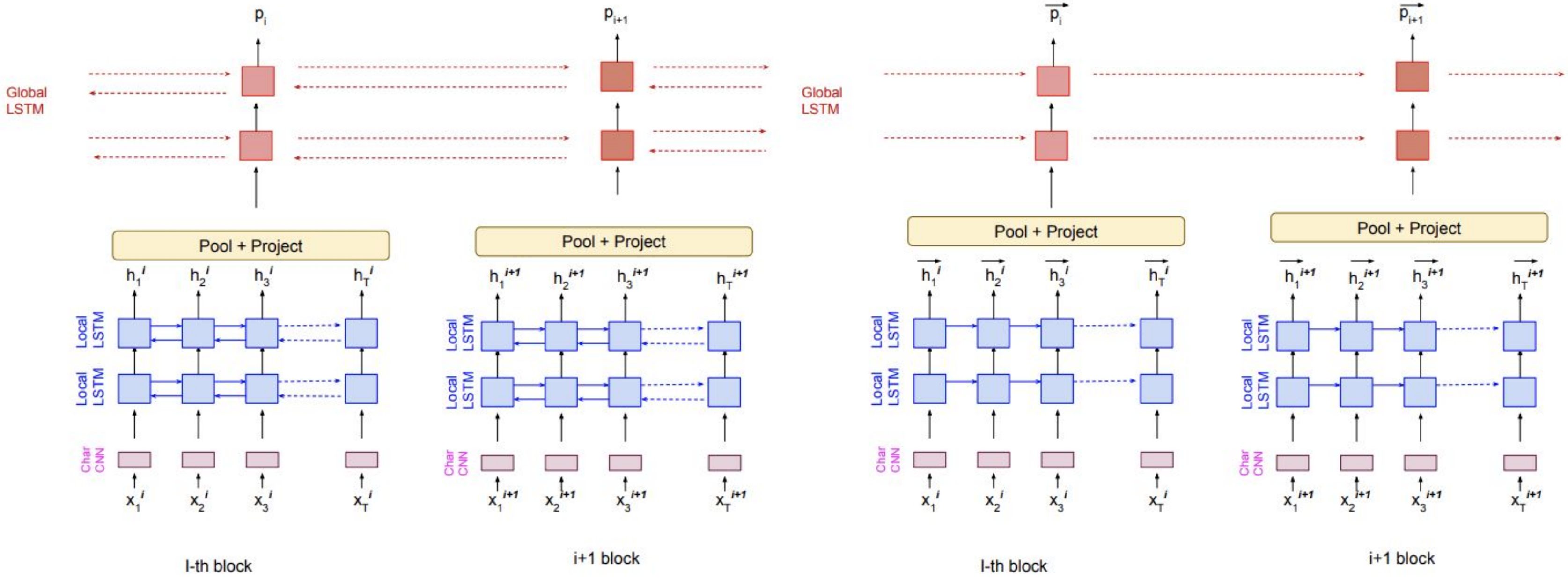
Jain et al. “SciREX: A challenging dataset for Document-level Information Extraction.” 2020.

Hierarchical Modeling

- Can be combined with a pre-training strategy

Hierarchical Modeling

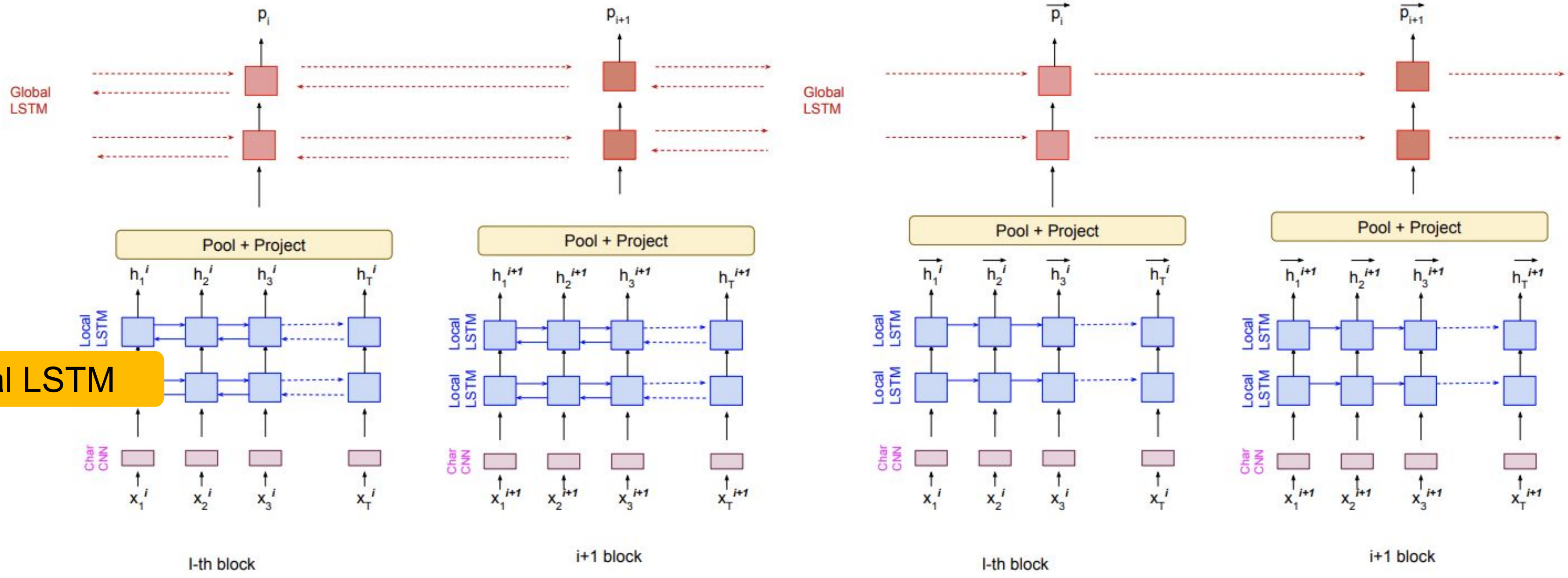
- Can be combined with a pre-training strategy



Chang et al. “Language Model Pre-training for Hierarchical Document Representation.” 2019.

Hierarchical Modeling

- Can be combined with a pre-training strategy

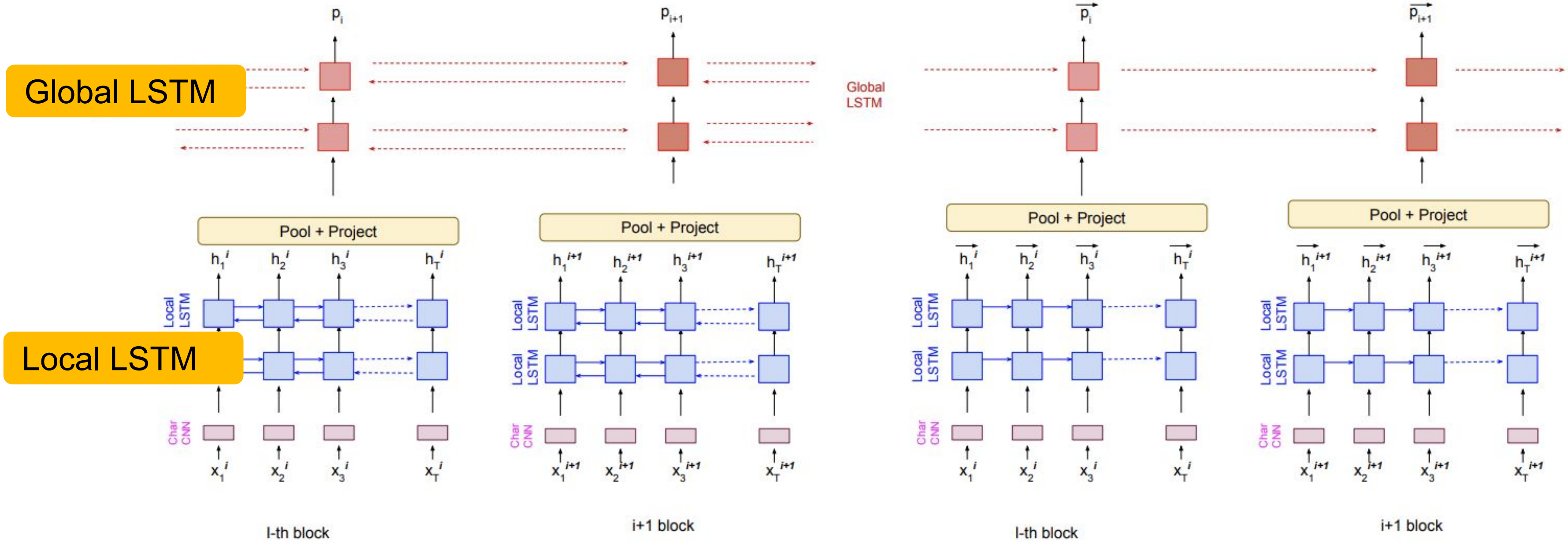


Local LSTM

Chang et al. “Language Model Pre-training for Hierarchical Document Representation.” 2019.

Hierarchical Modeling

- Can be combined with a pre-training strategy

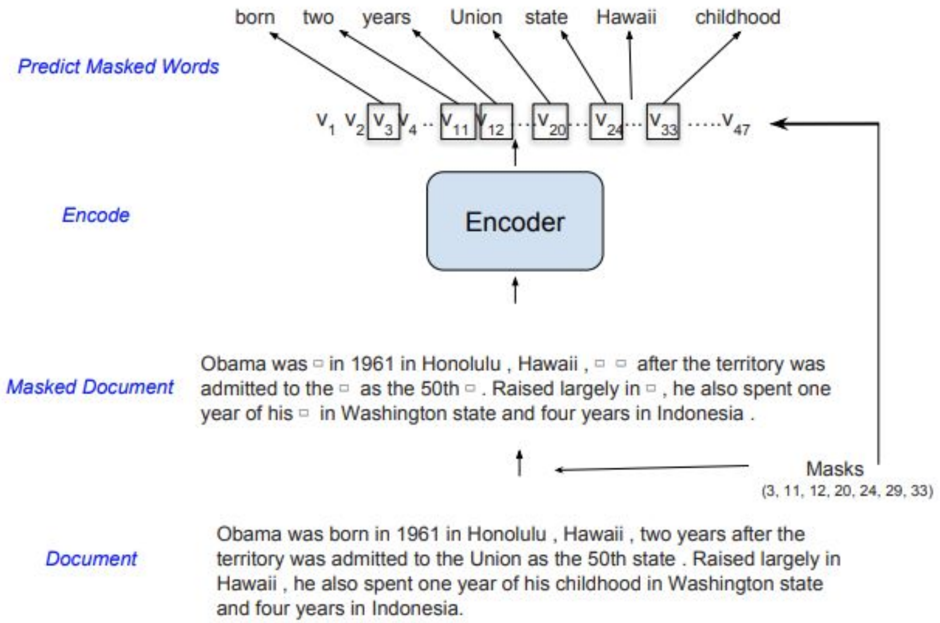


Chang et al. “Language Model Pre-training for Hierarchical Document Representation.” 2019.

Hierarchical Modeling

- Can be combined with a pre-training strategy

Masked LM



Chang et al. "Language Model Pre-training for Hierarchical Document Representation." 2019.

Hierarchical Modeling

- Can be combined with a pre-training strategy

Pre-training	MASK-LM BI-HLSTM	L+R-LM L+R-HLSTM
No	42.0 (0.0)	41.7 (0.0)
Local	50.6 (8.6)	48.4 (6.7)
Global	**51.8** (9.8)	**54.9** (13.2)
LSTM+ELMo$_{pool}$	44.6	
LSTM+Skip-Thought	46.0	

(a) Document Segmentation

Pre-training	MASK-LM BI-HLSTM	L+R-LM L+R-HLSTM
No	77.24 (0.0)	77.20 (0.0)
Local	79.17 (1.9)	78.36 (1.2)
Global	**79.92** (2.7)	**79.57** (2.4)
(Clark & Gardner, 2018)	73.31	

(b) Answer Passage Retrieval

(on TriviaQA-wiki)

Chang et al. “Language Model Pre-training for Hierarchical Document Representation.” 2019.

UWNLP AI2

Hierarchical Modeling

- Can be combined with a pre-training strategy

Pre-training	MASK-LM BI-HLSTM	L+R-LM L+R-HLSTM
No	42.0 (0.0)	41.7 (0.0)
Local	50.6 (8.6)	48.4 (6.7)
Global	**51.8** (9.8)	**54.9** (13.2)
LSTM+ELMo$_{pool}$	44.6	
LSTM+Skip-Thought	46.0	

(a) Document Segmentation

Pre-training	MASK-LM BI-HLSTM	L+R-LM L+R-HLSTM
No	77.24 (0.0)	77.20 (0.0)
Local	79.17 (1.9)	78.36 (1.2)
Global	**79.92** (2.7)	**79.57** (2.4)
(Clark & Gardner, 2018)	73.31	

(b) Answer Passage Retrieval

(on TriviaQA-wiki)

Chang et al. "Language Model Pre-training for Hierarchical Document Representation." 2019.

Hierarchical Modeling

- Can be combined with a pre-training strategy

Pre-training	MASK-LM BI-HLSTM	L+R-LM L+R-HLSTM
No	42.0 (0.0)	41.7 (0.0)
Local	50.6 (8.6)	48.4 (6.7)
Global	**51.8** (9.8)	**54.9** (13.2)
LSTM+ELMo$_{pool}$	44.6	
LSTM+Skip-Thought	46.0	

(a) Document Segmentation

Pre-training	MASK-LM BI-HLSTM	L+R-LM L+R-HLSTM
No	77.24 (0.0)	77.20 (0.0)
Local	79.17 (1.9)	78.36 (1.2)
Global	**79.92** (2.7)	**79.57** (2.4)
(Clark & Gardner, 2018)	73.31	

(b) Answer Passage Retrieval

(on TriviaQA-wiki)

Chang et al. "Language Model Pre-training for Hierarchical Document Representation." 2019.

Hierarchical Modeling

- Can be combined with a pre-training strategy

Pre-training	MASK-LM BI-HLSTM	L+R-LM L+R-HLSTM
No	42.0 (0.0)	41.7 (0.0)
Local	50.6 (8.6)	48.4 (6.7)
Global	**51.8** (9.8)	**54.9** (13.2)
LSTM+ELMo$_{pool}$	44.6	
LSTM+Skip-Thought	46.0	

(a) Document Segmentation

Pre-training	MASK-LM BI-HLSTM	L+R-LM L+R-HLSTM
No	77.24 (0.0)	77.20 (0.0)
Local	79.17 (1.9)	78.36 (1.2)
Global	**79.92** (2.7)	**79.57** (2.4)
(Clark & Gardner, 2018)	73.31	

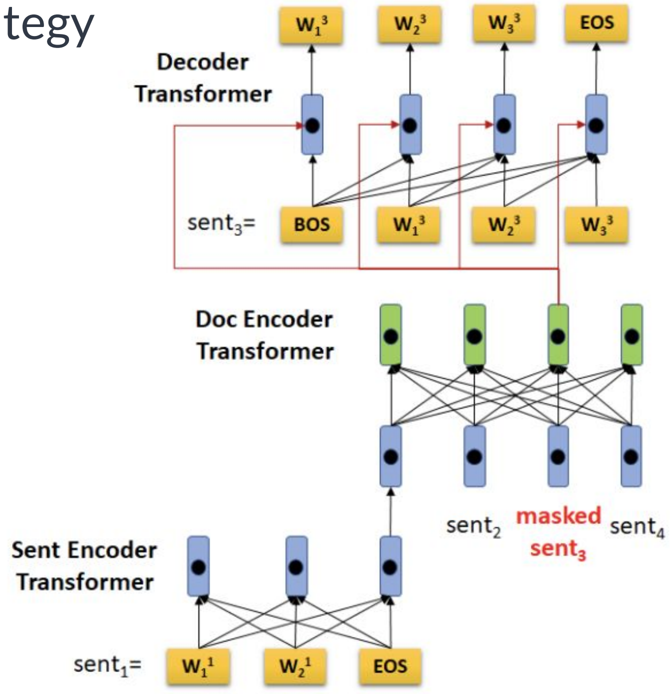
(b) Answer Passage Retrieval

(on TriviaQA-wiki)

Chang et al. "Language Model Pre-training for Hierarchical Document Representation." 2019.

Hierarchical Modeling

- Can be combined with a pre-training strategy

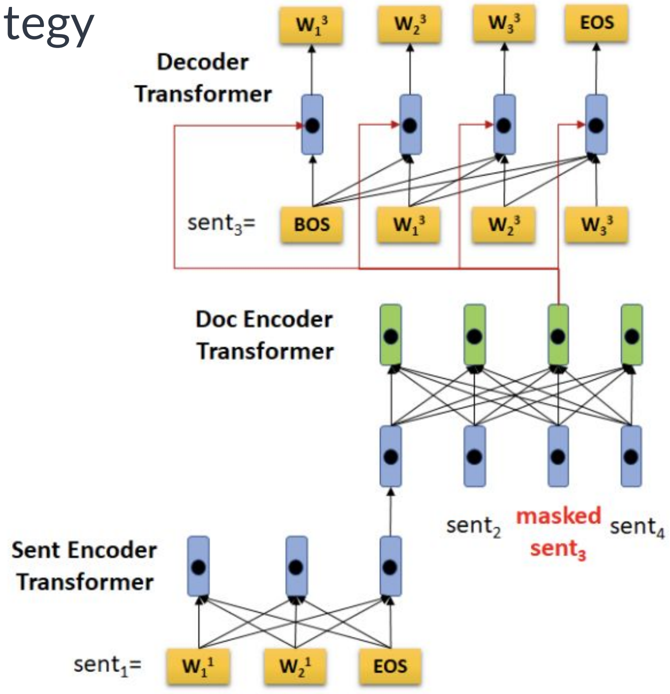


Zhang et al. “HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization.” 2019.

Hierarchical Modeling

- Can be combined with a pre-training strategy

Transformers instead of LSTMs



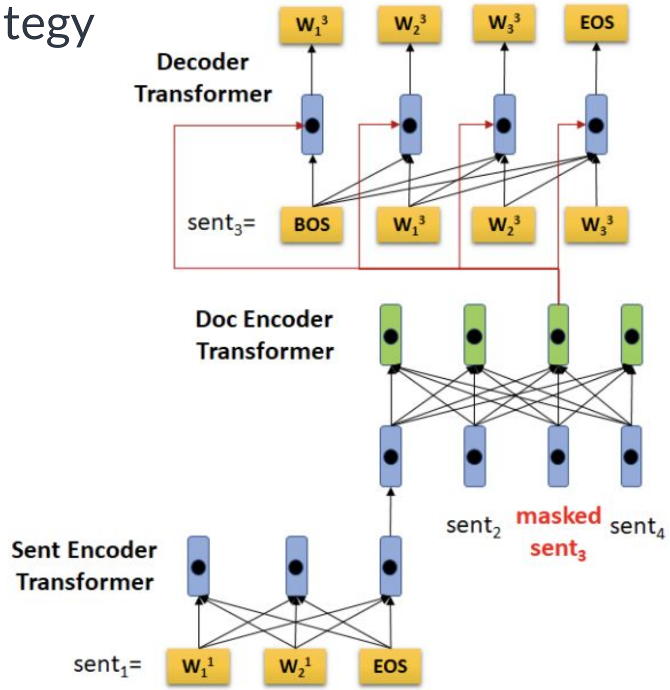
Zhang et al. "HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization." 2019.

Hierarchical Modeling

- Can be combined with a pre-training strategy

Transformers instead of LSTMs

Masked LM + Sentence Prediction



Zhang et al. “HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization.” 2019.

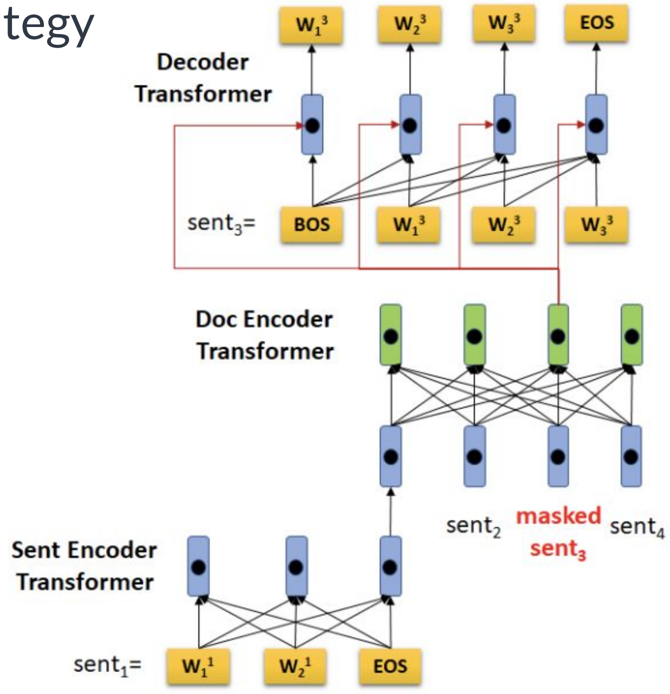
Hierarchical Modeling

- Can be combined with a pre-training strategy

Transformers instead of LSTMs

Masked LM + Sentence Prediction

Improves document summarization



Zhang et al. "HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization." 2019.

Finer-grained representation?

Finer-grained representation?

- Representation of one chunk depends on a chain of chunks.

```
This paper summarizes Category Cooccurrence Restrictions (CCRs)…

CCRs are Boolean conditions on the cooccurrence of categories in local
trees…

Their use leads to syntactic descriptions formulated entirely…
```

Relation: USED FOR

Example: Information Extraction

UWNLP AI2

Finer-grained representation?

- Representation of one chunk depends on a chain of chunks.

```
This paper summarizes Category Cooccurrence Restrictions (CCRs)…

CCRs are Boolean conditions on the cooccurrence of categories in local
trees…

Their use leads to syntactic descriptions formulated entirely…

      Relation: USED FOR
```

Example: Information Extraction

Finer-grained representation?

- Representation of one chunk depends on a chain of chunks.

`This paper summarizes` **Category Cooccurrence Restrictions (CCRs)**`…`

CCRs `are Boolean conditions on the cooccurrence of categories in local trees…`

Their `use leads to` syntactic descriptions `formulated entirely…`

`Relation: USED FOR`

Example: Information Extraction

Graph-based methods

- Representation of one chunk depends on a chain of chunks.

This paper summarizes Category Cooccurrence Restrictions (CCRs)…

CCRs are Boolean conditions on the cooccurrence
trees…

Can we update representations of chunks conditioned on other chunks multiple times?
→ **Graph propagation**

entirely…

Relation: USED FOR

Example: Information Extraction

Graph-based method in Coreference Resolution

- Iteratively refine span representations and antecedent representations

Representation of span i at iteration n $\boldsymbol{g}_i^n$

Update from iteration n to $n+1$ $\boldsymbol{g}_i^n \rightarrow \boldsymbol{g}_i^{n+1}$

Distributions of antecedents

$$P_n(y_i) = \frac{e^{s(\boldsymbol{g}_i^n, \boldsymbol{g}_{y_i}^n)}}{\sum_{y \in \boxed{\mathcal{Y}(i)}} e^{s(\boldsymbol{g}_i^n, \boldsymbol{g}_y^n))}}$$

A set of candidate antecedents

Lee et al. "Higher-order Coreference Resolution with Coarse-to-fine Inference." 2018.

UWNLP AI2

Graph-based method in Coreference Resolution

- Iteratively refine span representations and antecedent representations

Representation of span i at iteration n $\boldsymbol{g}_i^n$

Update from iteration n to $n+1$ $\boldsymbol{g}_i^n \rightarrow \boldsymbol{g}_i^{n+1}$

$$\boldsymbol{a}_i^n = \sum_{y_i \in \mathcal{Y}(i)} P_n(y_i) \cdot \boldsymbol{g}_{y_i}^n$$

$$\boldsymbol{f}_i^n = \sigma(\mathbf{W}_{\mathrm{f}}[\boldsymbol{g}_i^n, \boldsymbol{a}_i^n])$$

$$\boldsymbol{g}_i^{n+1} = \boldsymbol{f}_i^n \circ \boldsymbol{g}_i^n + (\mathbf{1} - \boldsymbol{f}_i^n) \circ \boldsymbol{a}_i^n$$

Distributions of antecedents

$$P_n(y_i) = \frac{e^{s(\boldsymbol{g}_i^n, \boldsymbol{g}_{y_i}^n)}}{\sum_{y \in \boxed{\mathcal{Y}(i)}} e^{s(\boldsymbol{g}_i^n, \boldsymbol{g}_y^n))}}$$

A set of candidate antecedents

Lee et al. "Higher-order Coreference Resolution with Coarse-to-fine Inference." 2018.

UWNLP AI2

Graph-based method in Coreference Resolution

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Lee et al. (2017)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
+ ELMo (Peters et al., 2018)	80.1	77.2	78.6	69.8	66.5	68.1	66.4	62.9	64.6	70.4
+ hyperparameter tuning	80.7	78.8	79.8	71.7	68.7	70.2	67.2	66.8	67.0	72.3
+ coarse-to-fine inference	80.4	**79.9**	80.1	71.0	**70.0**	70.5	67.5	**67.2**	67.3	72.6
+ second-order inference	**81.4**	79.5	**80.4**	**72.2**	69.5	**70.8**	**68.2**	67.1	**67.6**	**73.0**

Lee et al. "Higher-order Coreference Resolution with Coarse-to-fine Inference." 2018.

Graph-based method in Coreference Resolution

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Lee et al. (2017)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
+ ELMo (Peters et al., 2018)	80.1	77.2	78.6	69.8	66.5	68.1	66.4	62.9	64.6	70.4
+ hyperparameter tuning	80.7	78.8	79.8	71.7	68.7	70.2	67.2	66.8	67.0	72.3
+ coarse-to-fine inference	80.4	**79.9**	80.1	71.0	**70.0**	70.5	67.5	**67.2**	67.3	72.6
+ second-order inference	**81.4**	79.5	**80.4**	**72.2**	69.5	**70.8**	**68.2**	67.1	**67.6**	**73.0**

Establish a standard in coference resolution

Lee et al. “Higher-order Coreference Resolution with Coarse-to-fine Inference.” 2018.

UWNLP AI2

Graph-based method for Information Extraction

Idea: Information Extraction as *Span Classification*

```
This paper summarizes Category Cooccurrence Restrictions (CCRs)…

CCRs are Boolean conditions on the cooccurrence of categories in local trees…

Their use leads to syntactic descriptions formulated entirely…
```

Graph-based method for Information Extraction

Idea: Information Extraction as *Span Classification*

This paper summarizes **Category Cooccurrence Restrictions (CCRs)**…

CCRs are Boolean conditions on the cooccurrence of categories in local trees…

Their use leads to syntactic descriptions formulated entirely…

Feedforward(**)** = **Category Cooccurrence Restrictions**

CCRs

Graph-based method for Information Extraction

Idea: Information Extraction as *Span Classification*

```
This paper summarizes Category Cooccurrence Restrictions (CCRs)…

CCRs are Boolean conditions on the cooccurrence of categories in local trees…

Their use leads to syntactic descriptions formulated entirely…
```

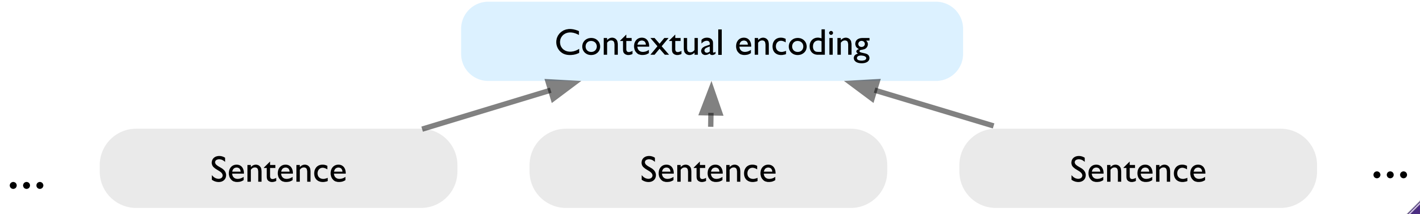
Feedforward(CCRs **)** = **Category Cooccurrence Restrictions**

Feedforward(Their **,** Syntactic description **)** = **USED FOR**

DyGIE++

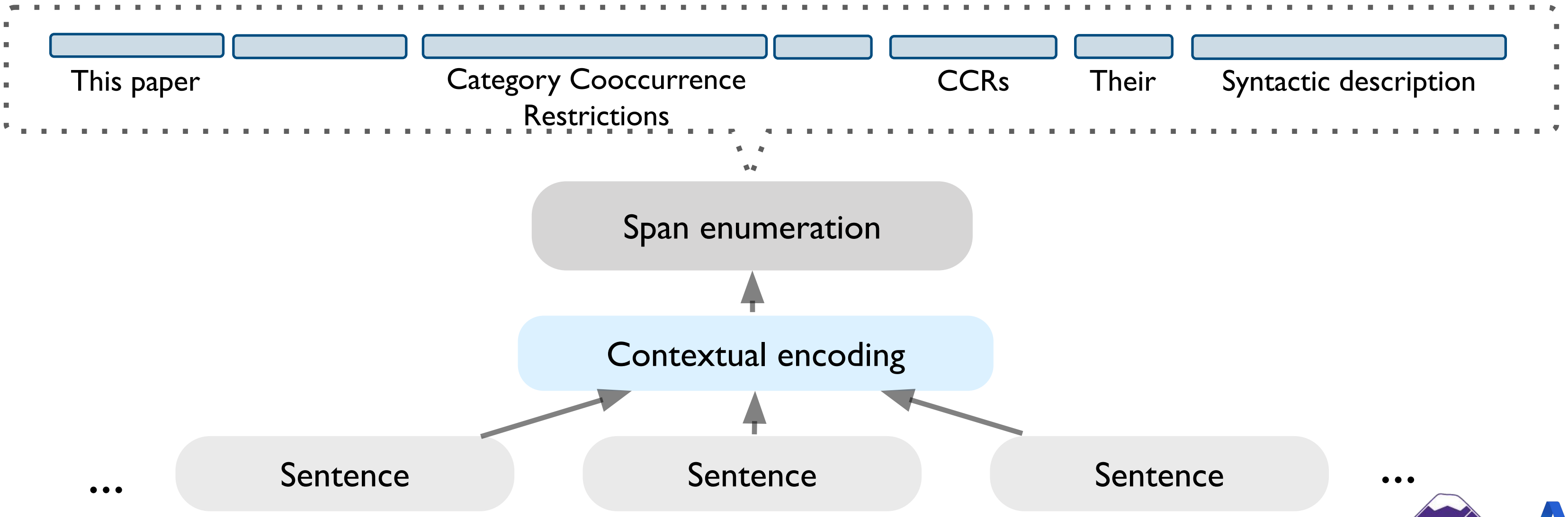
Wadden et al. “Entity, Relations, and Event Extraction with Contextualized Span Representations.” 2019.

DyGIE++



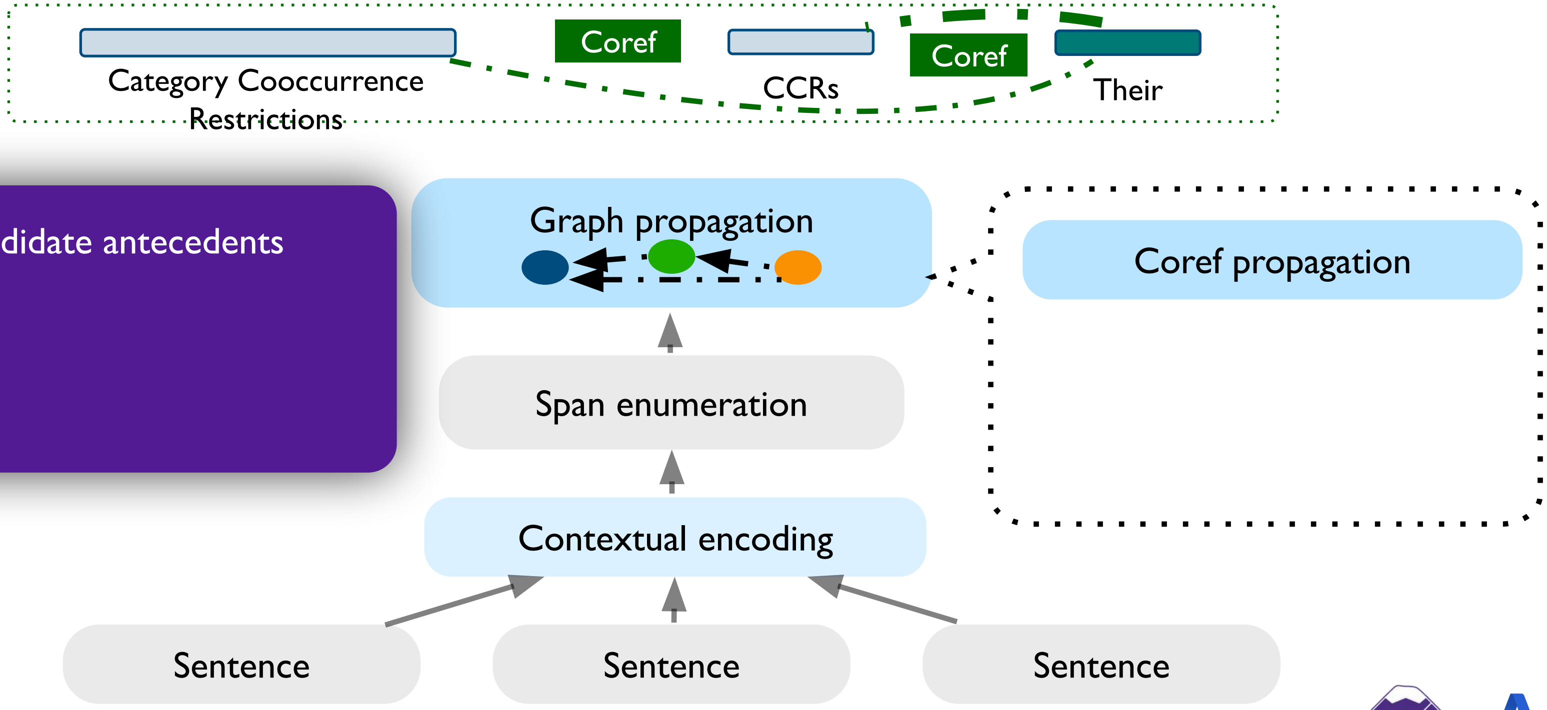
Wadden et al. “Entity, Relations, and Event Extraction with Contextualized Span Representations.” 2019.

DyGIE++



Wadden et al. “Entity, Relations, and Event Extraction with Contextualized Span Representations.” 2019.

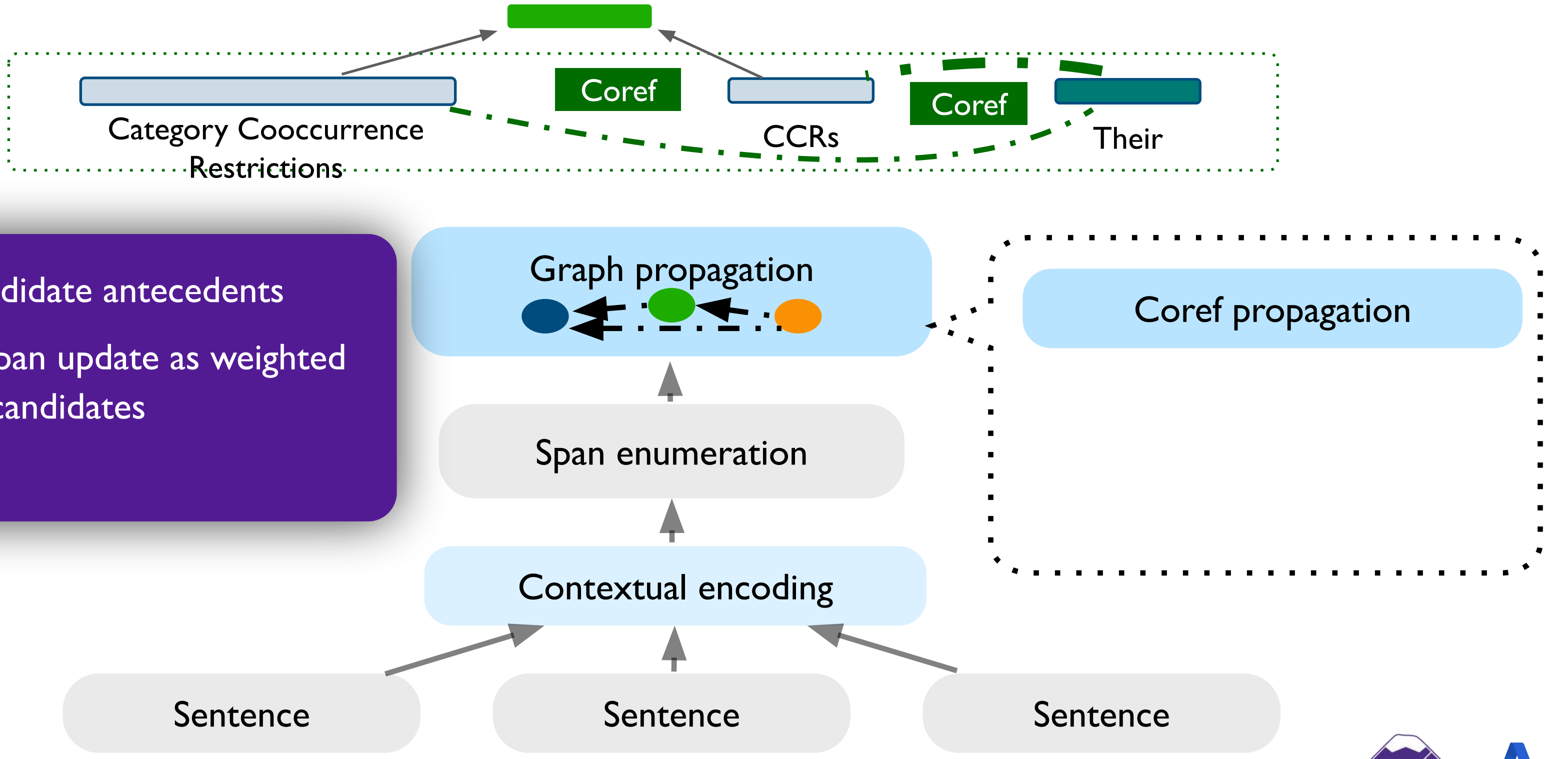
DyGIE++



- Identify candidate antecedents

Wadden et al. "Entity, Relations, and Event Extraction with Contextualized Span Representations." 2019.

DyGIE++

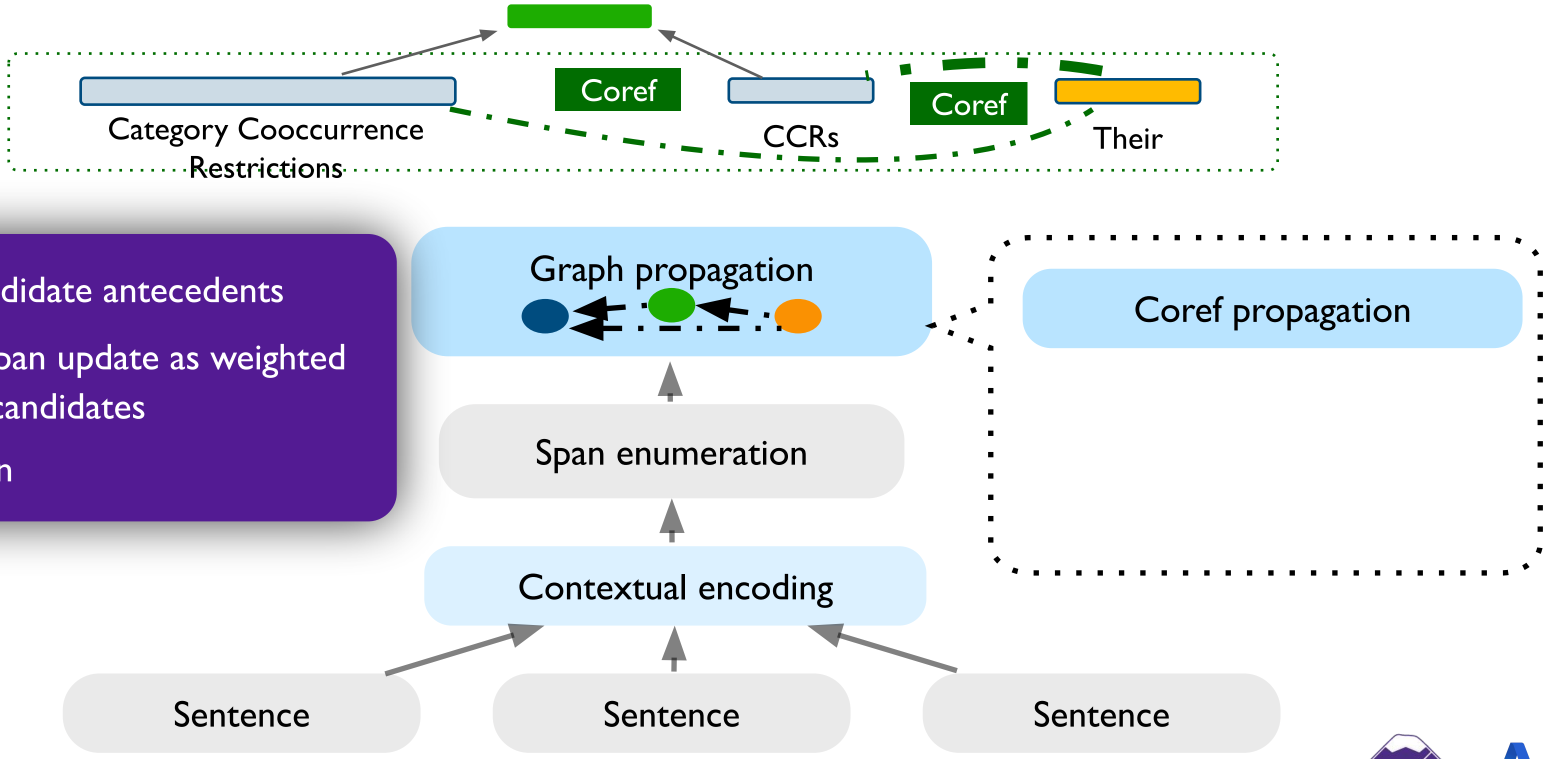


- Identify candidate antecedents
- Compute span update as weighted average of candidates

Wadden et al. "Entity, Relations, and Event Extraction with Contextualized Span Representations." 2019.

UWNLP AI2

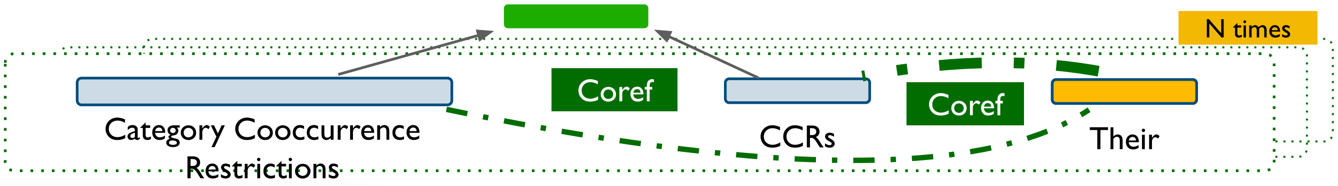
DyGIE++



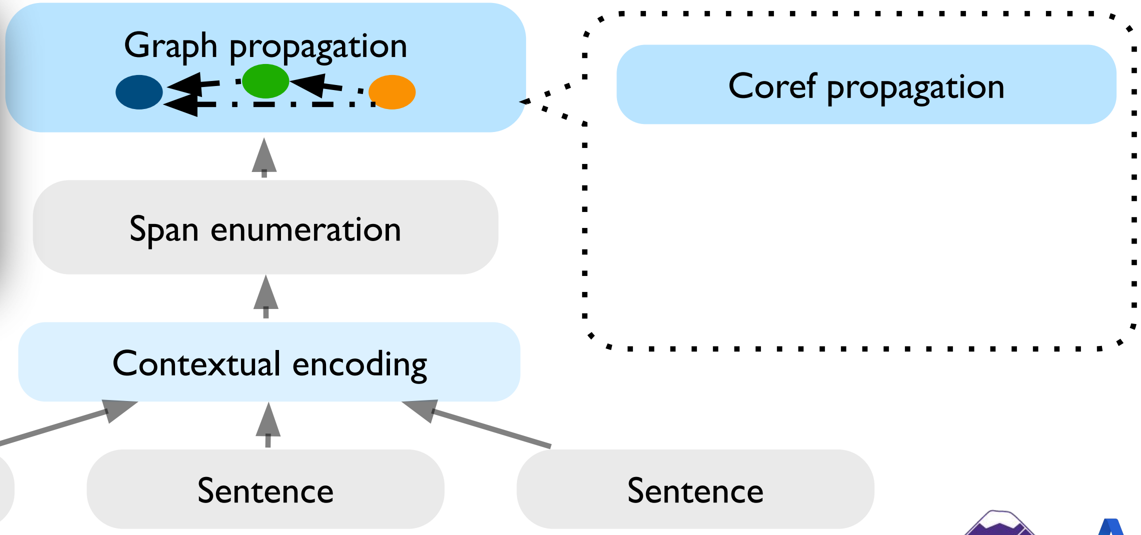
- Identify candidate antecedents
- Compute span update as weighted average of candidates
- Update span

Wadden et al. "Entity, Relations, and Event Extraction with Contextualized Span Representations." 2019.

DyGIE++



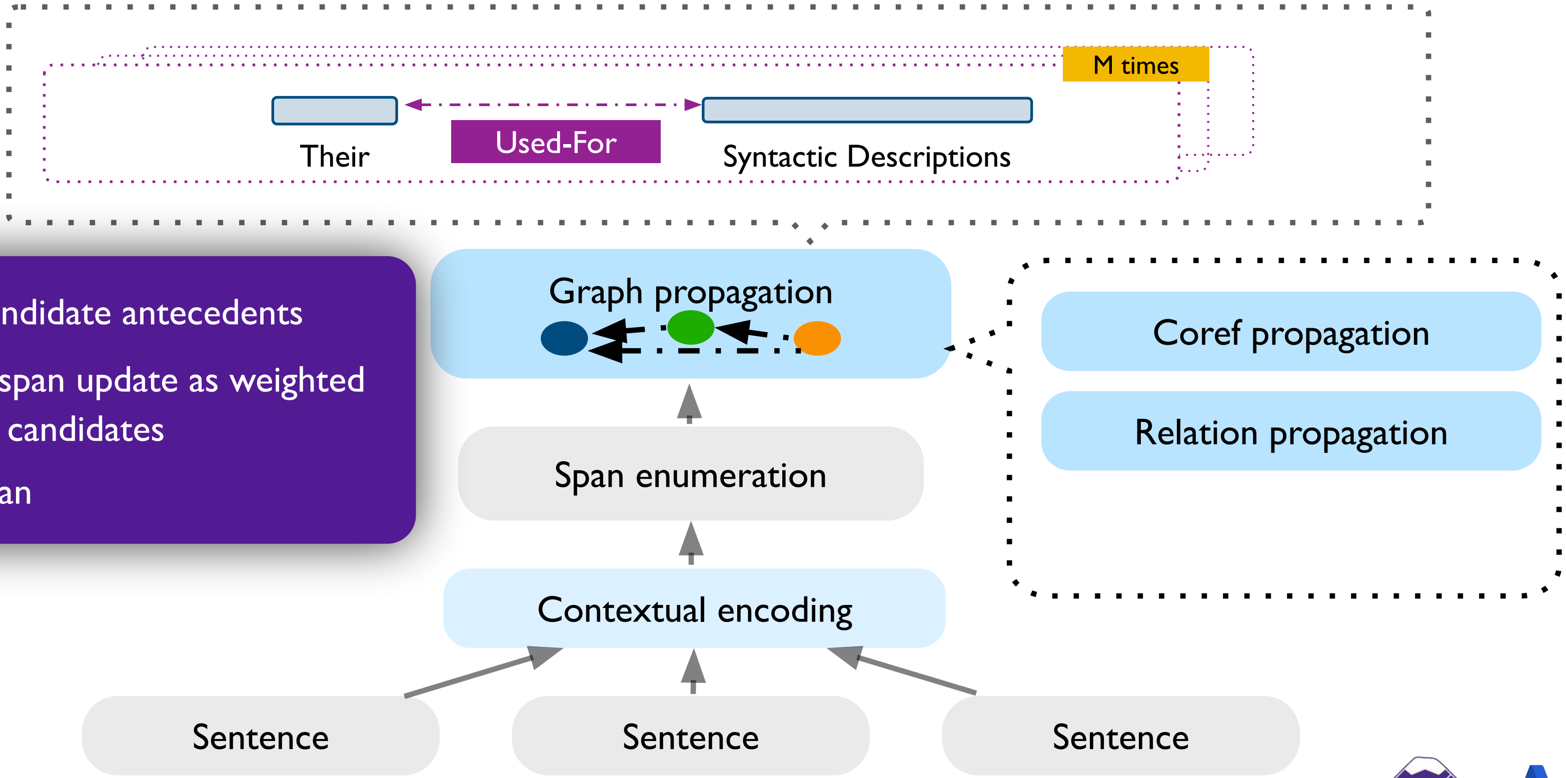
- Identify candidate antecedents
- Compute span update as weighted average of candidates
- Update span



Wadden et al. “Entity, Relations, and Event Extraction with Contextualized Span Representations.” 2019.

DyGIE++

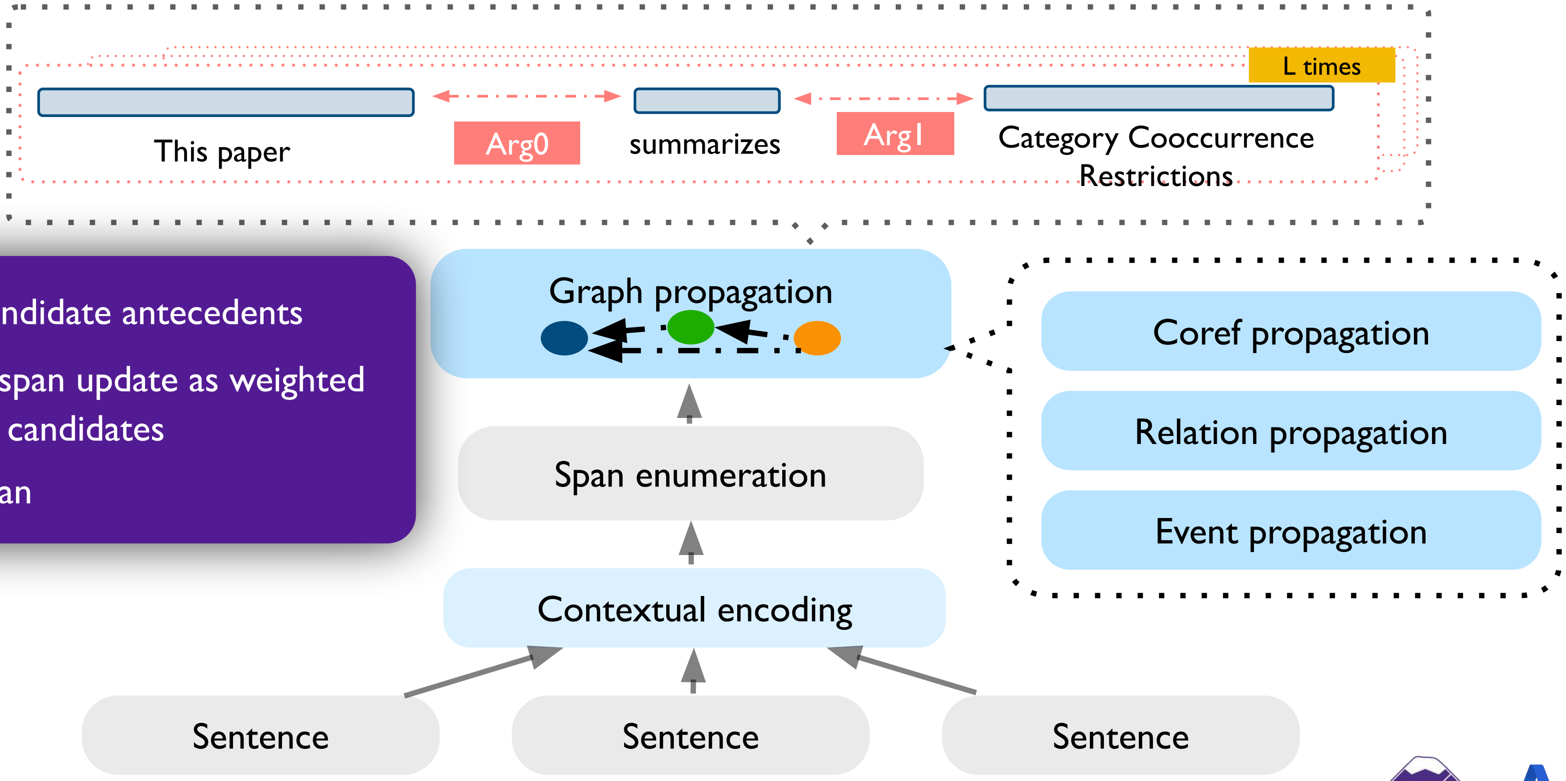
- Identify candidate antecedents
- Compute span update as weighted average of candidates
- Update span



Wadden et al. “Entity, Relations, and Event Extraction with Contextualized Span Representations.” 2019.

DyGIE++

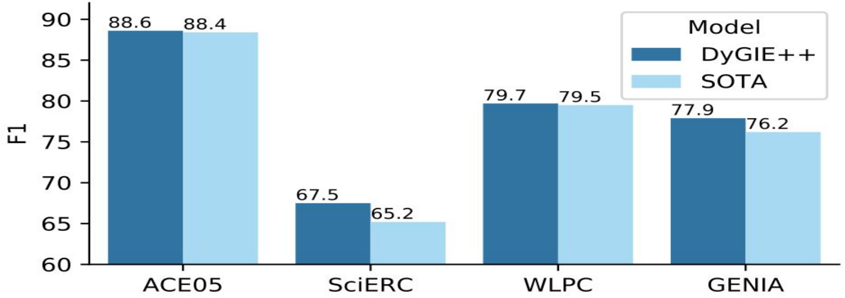
- Identify candidate antecedents
- Compute span update as weighted average of candidates
- Update span



Wadden et al. “Entity, Relations, and Event Extraction with Contextualized Span Representations.” 2019.

DyGIE++

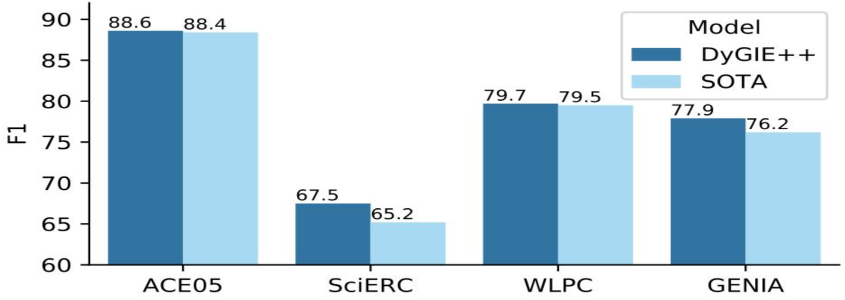
- SOTA results on multiple information extraction tasks



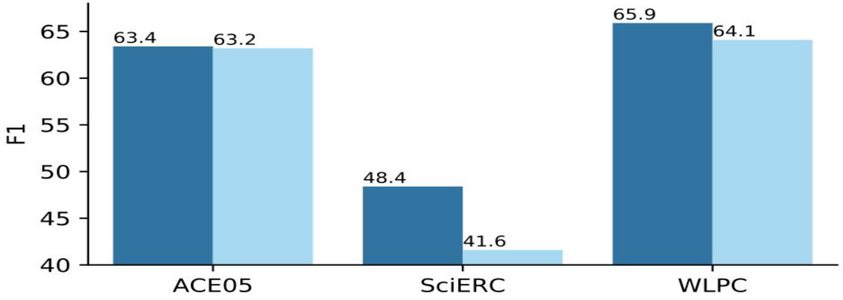
Named entity recognition

DyGIE++

- SOTA results on multiple information extraction tasks



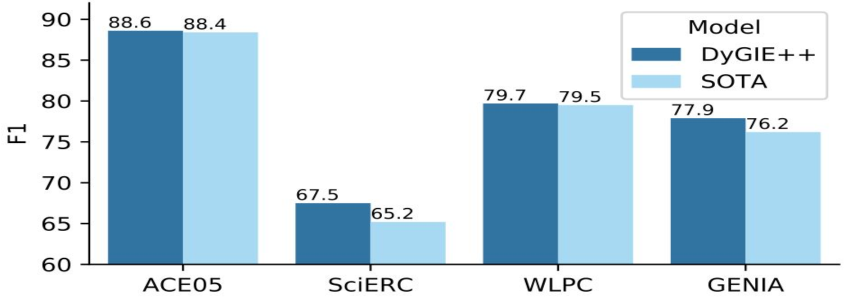
Named entity recognition



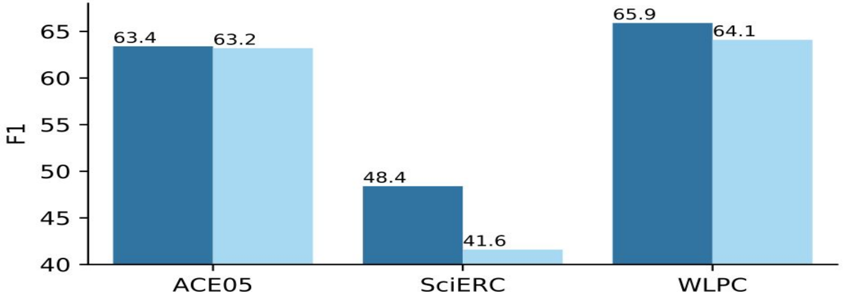
Relation extraction

DyGIE++

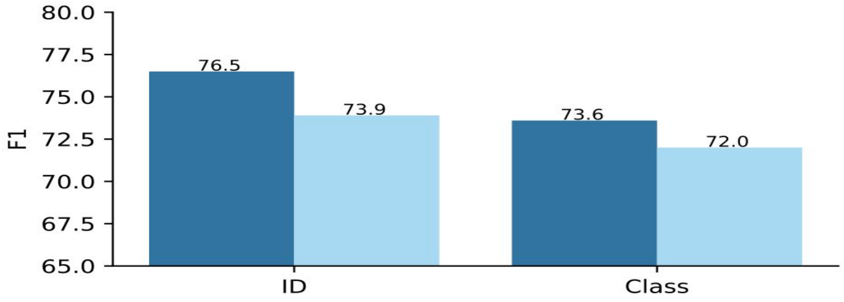
- SOTA results on multiple information extraction tasks



Named entity recognition



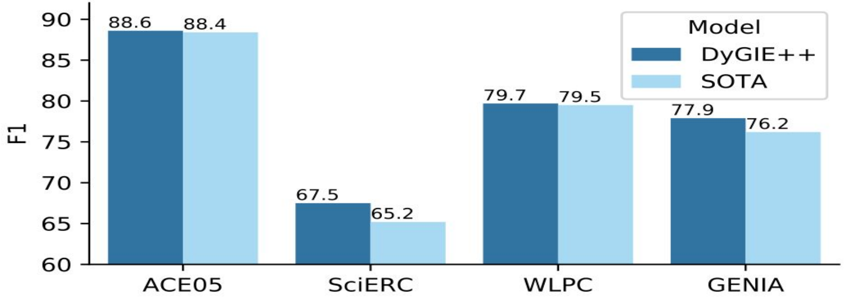
Relation extraction



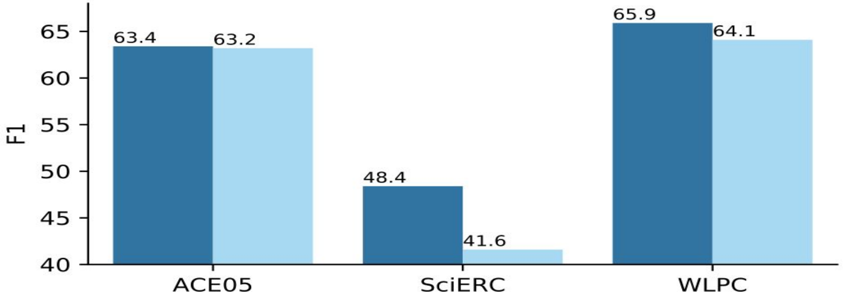
Event extraction

DyGIE++

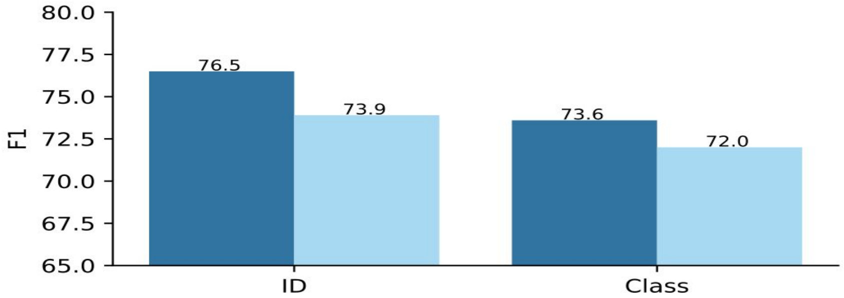
- SOTA results on multiple information extraction tasks



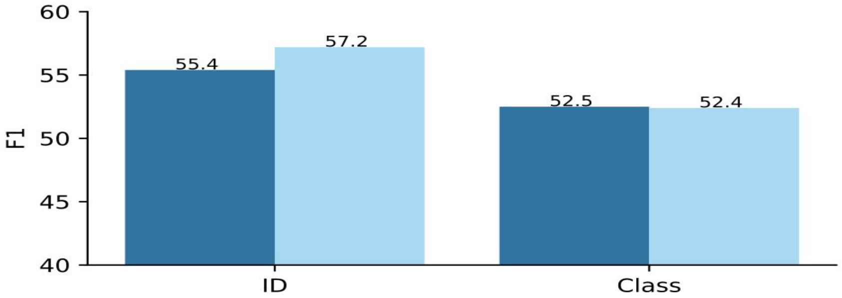
Named entity recognition



Relation extraction



Event extraction

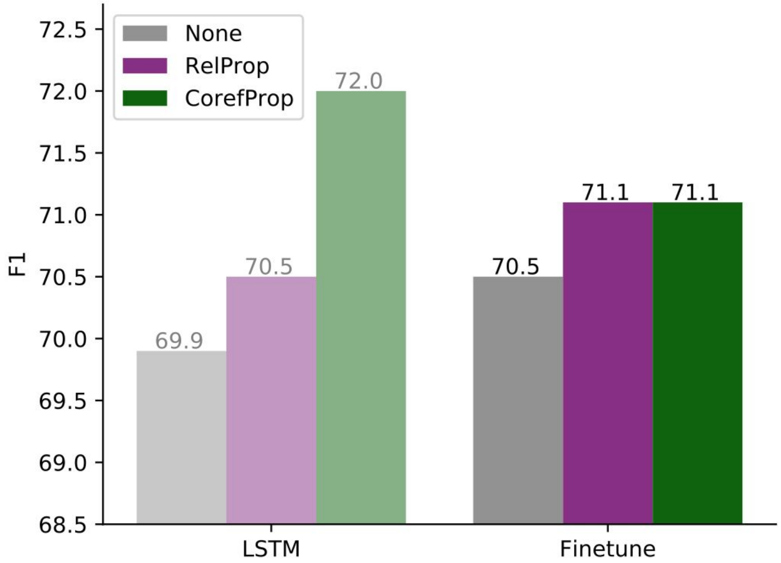


Event argument classification

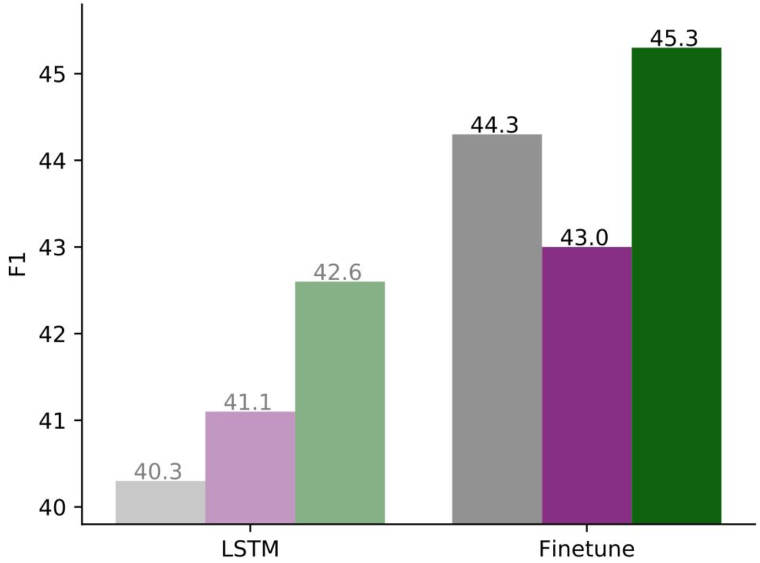
DyGIE++

- SOTA results on multiple information extraction tasks

Named entity recognition



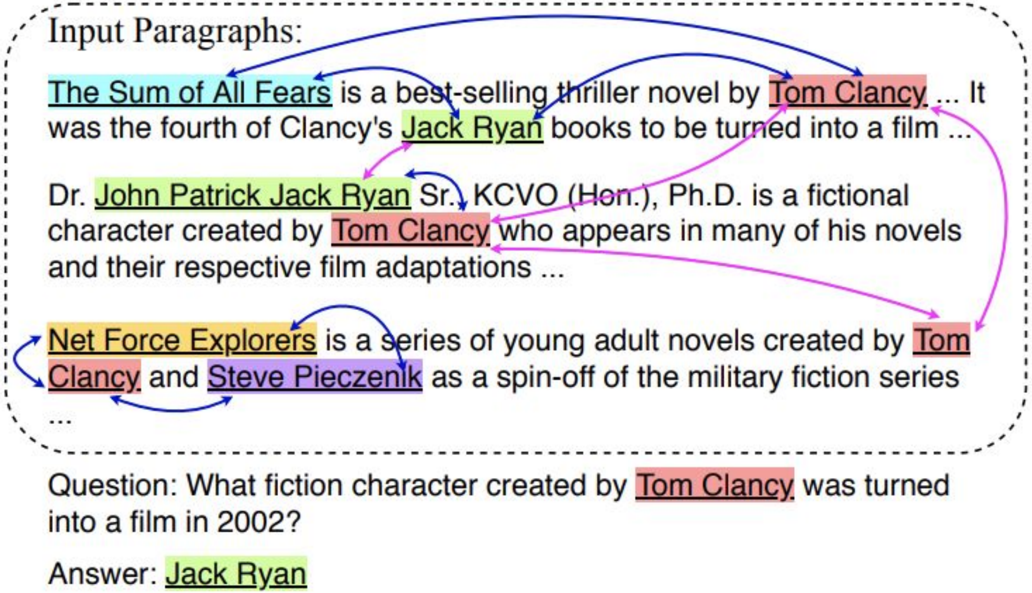
Relation extraction



Wadden et al. “Entity, Relations, and Event Extraction with Contextualized Span Representations.” 2019.

Graph-based Methods

- Also became standard approach in multi-hop question answering
  - Song et al. 2018: uses graph-structured passage representations / Graph Neural Networks
  - Xiao et al. 2019: proposes dynamically fused graph network



Graph-based Methods w/ external knowledge

- Can we exploit external knowledge for better modeling of long documents?

Fang et al. "Hierarchical Graph Network for Multi-hop Question Answering." 2020.

Graph-based Methods w/ external knowledge

- Can we exploit external knowledge for better modeling of long documents?

Question: The director of the romantic comedy "**Big Stone Gap**" is based in what New York city?

Article: **Big Stone Gap**

Big Stone Gap is a 2014 American drama romantic comedy film written and directed by **Adriana Trigiani** and produced by (...)

Article: **Adriana Trigiani**

Adriana Trigiani is an Italian American best-selling author (...) based in **Greenwich Village, New York City**.

Answer: **Greenwich Village, New York City**

Fang et al. "Hierarchical Graph Network for Multi-hop Question Answering." 2020.

Graph-based Methods w/ external knowledge

- Can we exploit external knowledge for better modeling of long documents?

Question: The director of the romantic comedy "**Big Stone Gap**" is based in what New York city?

Article: **Big Stone Gap**

Big Stone Gap is a 2014 American drama romantic comedy film written and directed by **Adriana Trigiani** and produced by (...)

Hyperlink

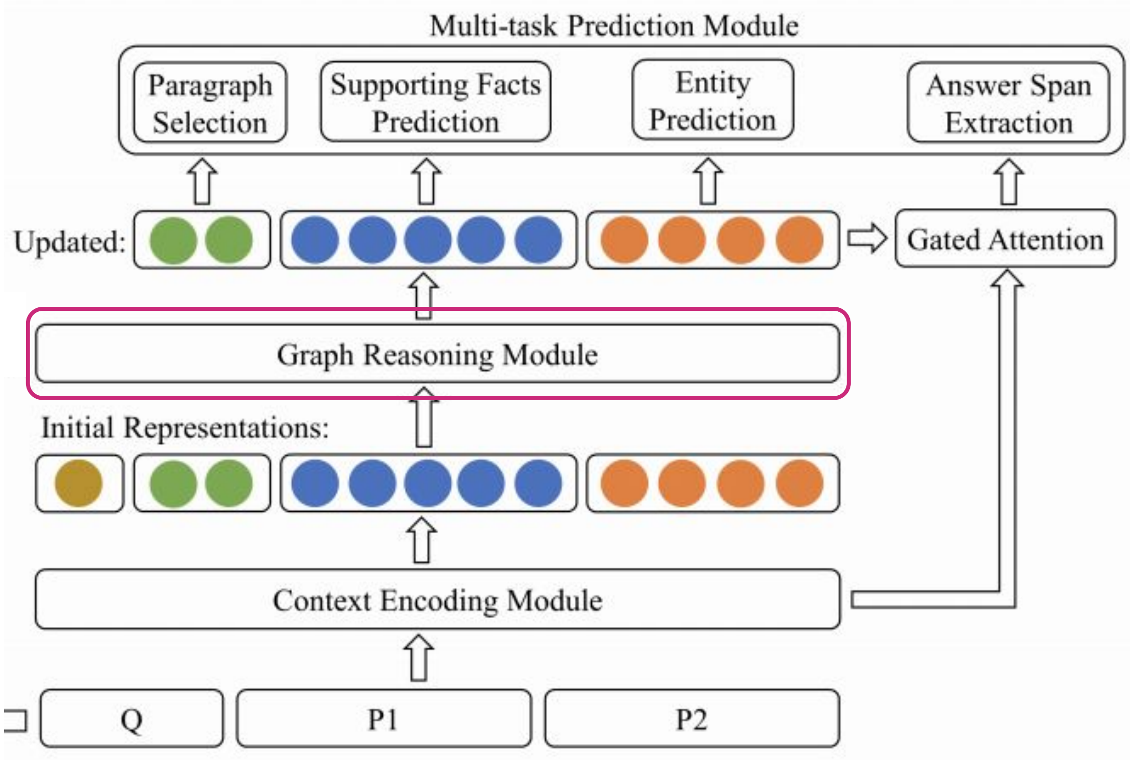
Article: **Adriana Trigiani**

Adriana Trigiani is an American best-selling author (...) based in **Greenwich Village, New York City**.

Answer: **Greenwich Village, New York City**

Fang et al. "Hierarchical Graph Network for Multi-hop Question Answering." 2020.

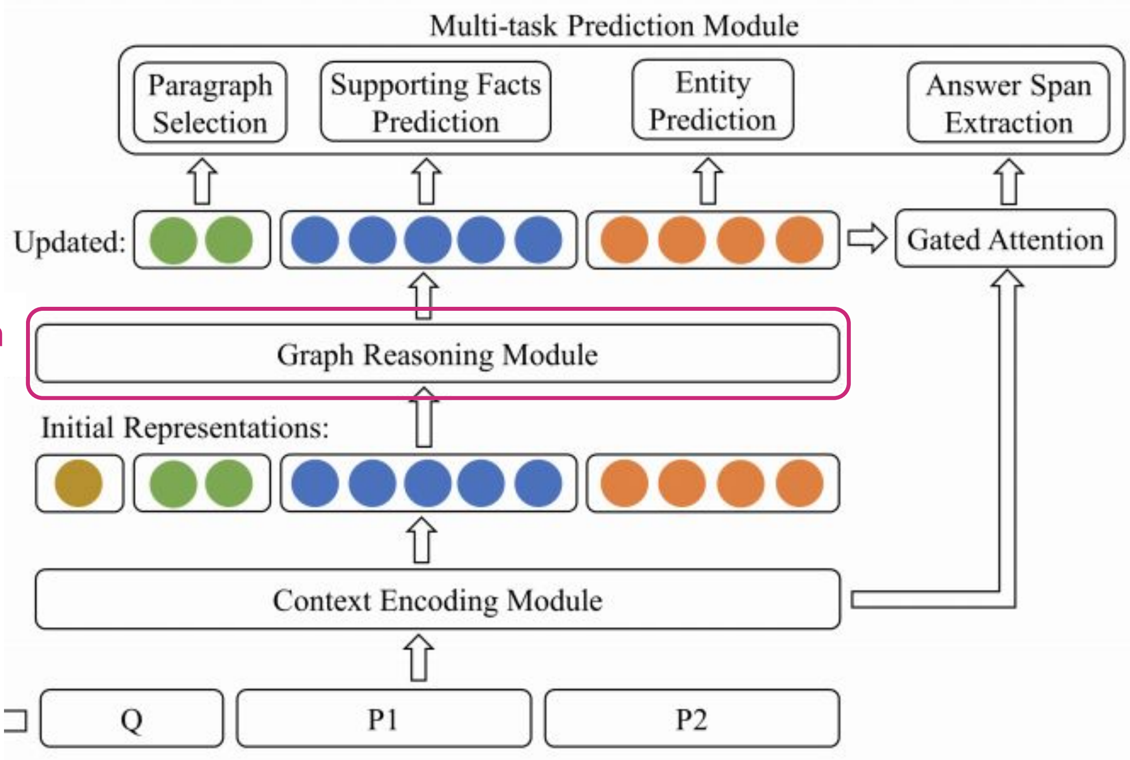
Hierarchical Graph Network



Fang et al. "Hierarchical Graph Network for Multi-hop Question Answering." 2020.

Hierarchical Graph Network

1) Graph Construction



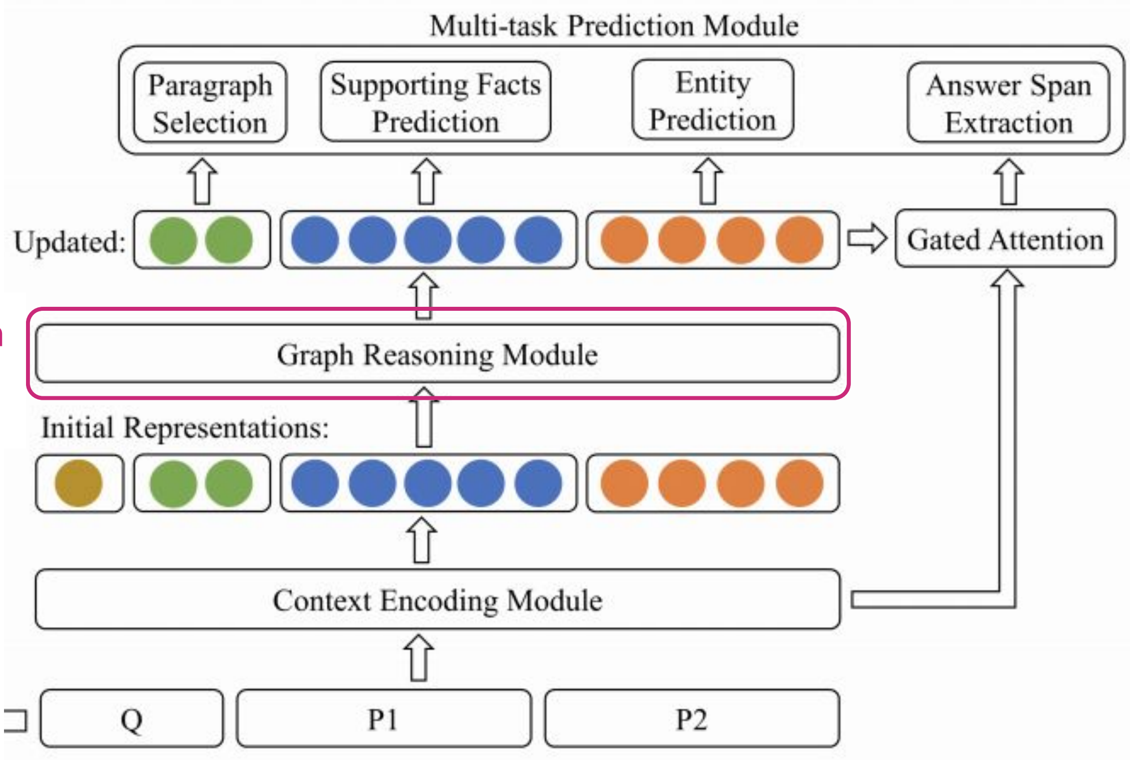
Fang et al. "Hierarchical Graph Network for Multi-hop Question Answering." 2020.

UWNLP Ai2

Hierarchical Graph Network

1) Graph Construction

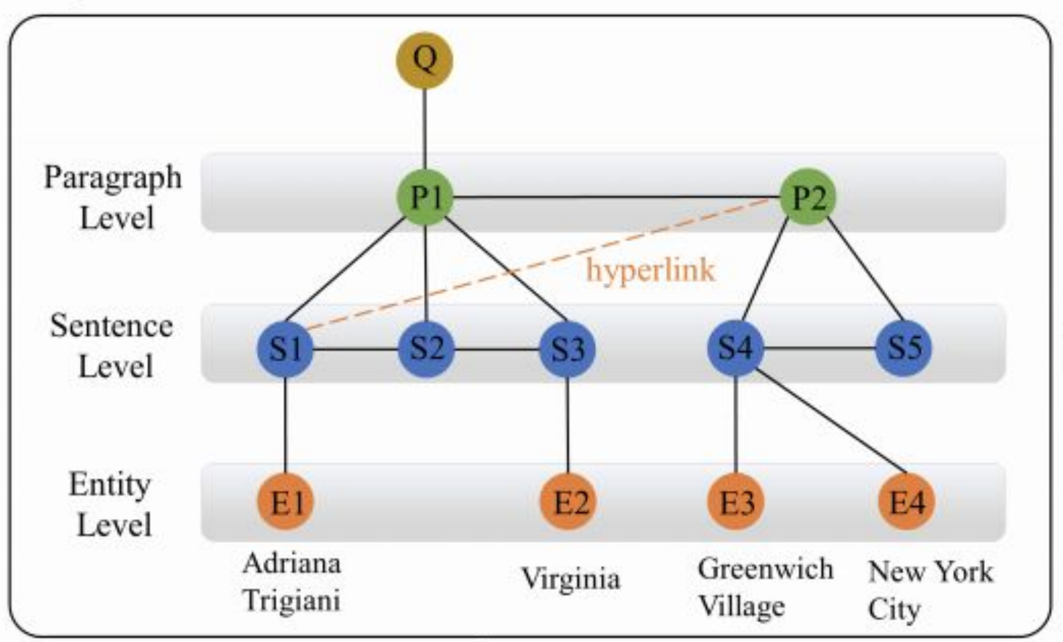
2) Graph Propagation



Fang et al. “Hierarchical Graph Network for Multi-hop Question Answering.” 2020.

Hierarchical Graph Network

1. Graph Construction

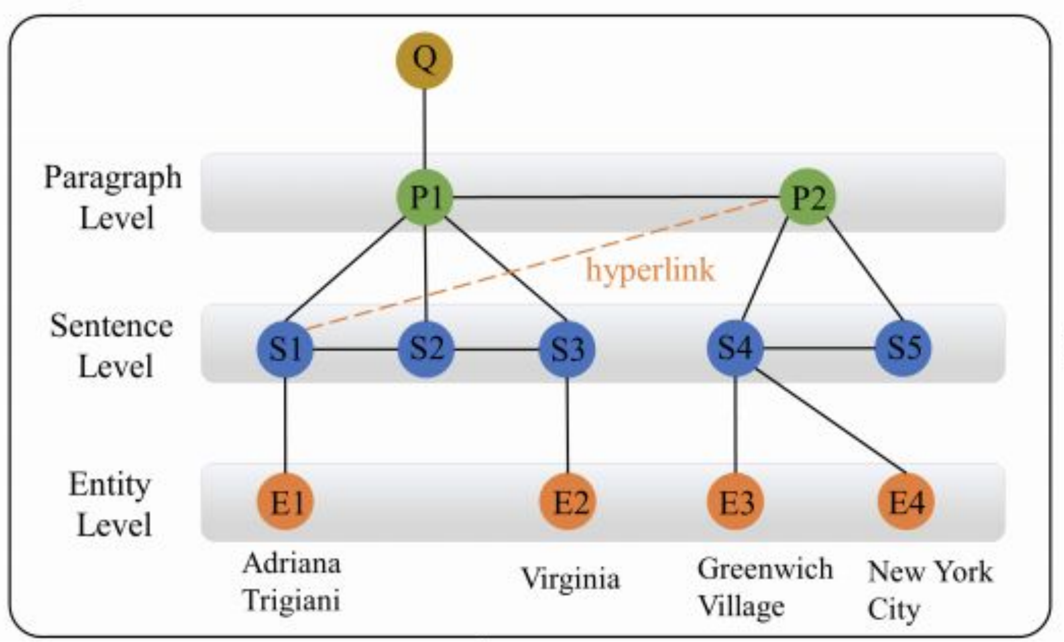


- Hierarchical structure
  - Entity-Sentence
  - Sentence-Paragraph
  - Sentence-Sentence
  - Paragraph-Paragraph

Fang et al. “Hierarchical Graph Network for Multi-hop Question Answering.” 2020.

Hierarchical Graph Network

1. Graph Construction



- Hierarchical structure
  - Entity-Sentence
  - Sentence-Paragraph
  - Sentence-Sentence
  - Paragraph-Paragraph
- From hyperlink
  - Sentence-paragraph

Fang et al. “Hierarchical Graph Network for Multi-hop Question Answering.” 2020.

Hierarchical Graph Network

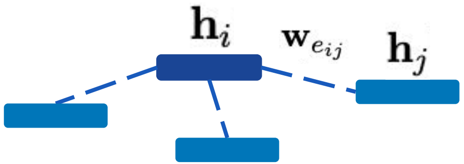
2. Graph Propagation

$\mathbf{h}_i$: a representation of a vertex *i* (either question, entity, sentence, paragraph)

Hierarchical Graph Network

2. Graph Propagation

$\mathbf{h}_i$: a representation of a vertex *i* (either question, entity, sentence, paragraph)



$$\rightarrow f([\mathbf{h}_i; \mathbf{h}_j]\mathbf{w}_{e_{ij}})$$

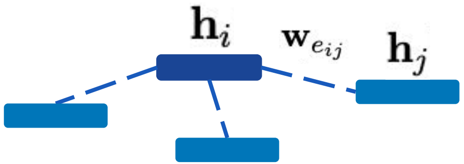
($\mathbf{w}_{e_{ij}}$ is a vector for an edge between vertex *i* and vertex *j*)

Fang et al. “Hierarchical Graph Network for Multi-hop Question Answering.” 2020.

Hierarchical Graph Network

2. Graph Propagation

$\mathbf{h}_i$: a representation of a vertex *i* (either question, entity, sentence, paragraph)



$$\rightarrow \quad f([\mathbf{h}_i; \mathbf{h}_j]\mathbf{w}_{e_{ij}}) \quad \rightarrow \quad \alpha_{ij} = \frac{\exp(f([\mathbf{h}_i; \mathbf{h}_j]\mathbf{w}_{e_{ij}}))}{\sum_{k \in \mathcal{N}_i} \exp(f([\mathbf{h}_i; \mathbf{h}_k]\mathbf{w}_{e_{ik}}))}$$

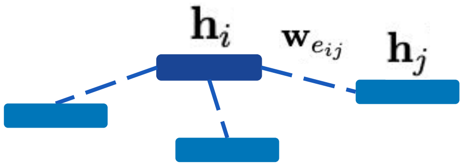
($\mathbf{w}_{e_{ij}}$ is a vector for an edge between vertex *i* and vertex *j*)

Fang et al. "Hierarchical Graph Network for Multi-hop Question Answering." 2020.

Hierarchical Graph Network

2. Graph Propagation

$\mathbf{h}_i$: a representation of a vertex *i* (either question, entity, sentence, paragraph)



$$f([\mathbf{h}_i; \mathbf{h}_j]\mathbf{w}_{e_{ij}})$$

$$\alpha_{ij} = \frac{\exp(f([\mathbf{h}_i; \mathbf{h}_j]\mathbf{w}_{e_{ij}}))}{\sum_{k \in \mathcal{N}_i} \exp(f([\mathbf{h}_i; \mathbf{h}_k]\mathbf{w}_{e_{ik}}))}$$

($\mathbf{w}_{e_{ij}}$ is a vector for an edge between vertex *i* and vertex *j*)

$$\mathbf{h}'_i = \text{LeakyRelu}\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij}\mathbf{h}_j\mathbf{W}\Big)$$

updated representation of a vertex *i*

Fang et al. “Hierarchical Graph Network for Multi-hop Question Answering.” 2020.

Hierarchical Graph Network

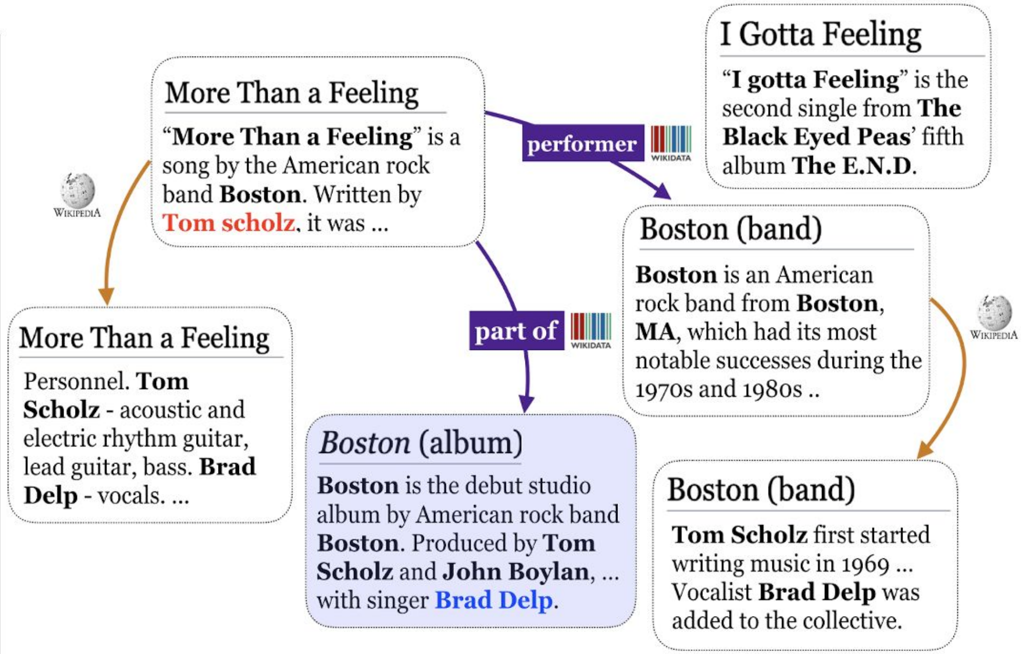
- SOTA on HotpotQA + ablation showing that graph reasoning is the key

	Q-P, P-S	Q-E, P-E	P-P, S-S	Ans F1	Sup F1	Joint F1
No Graph				80.6	85.8	71.0
Graph	✔			81.7	88.4	73.8
	✔	✔		82.1	88.4	74.1
	✔	✔	✔	**82.2**	**88.6**	**74.4**

Fang et al. "Hierarchical Graph Network for Multi-hop Question Answering." 2020.

Graph-based Methods w/ external knowledge

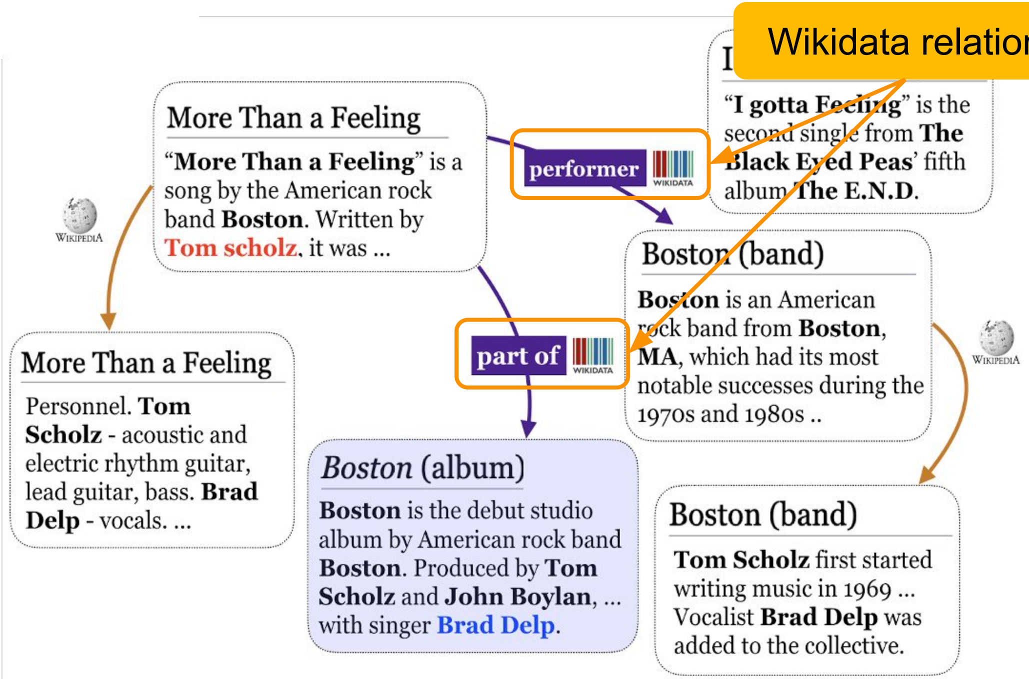
Question: Who sang More than a Feeling by Boston?



Min et al. "Knowledge-guided Text Retrieval and Reading for Open Domain Question Answering." 2019.

Graph-based Methods w/ external knowledge

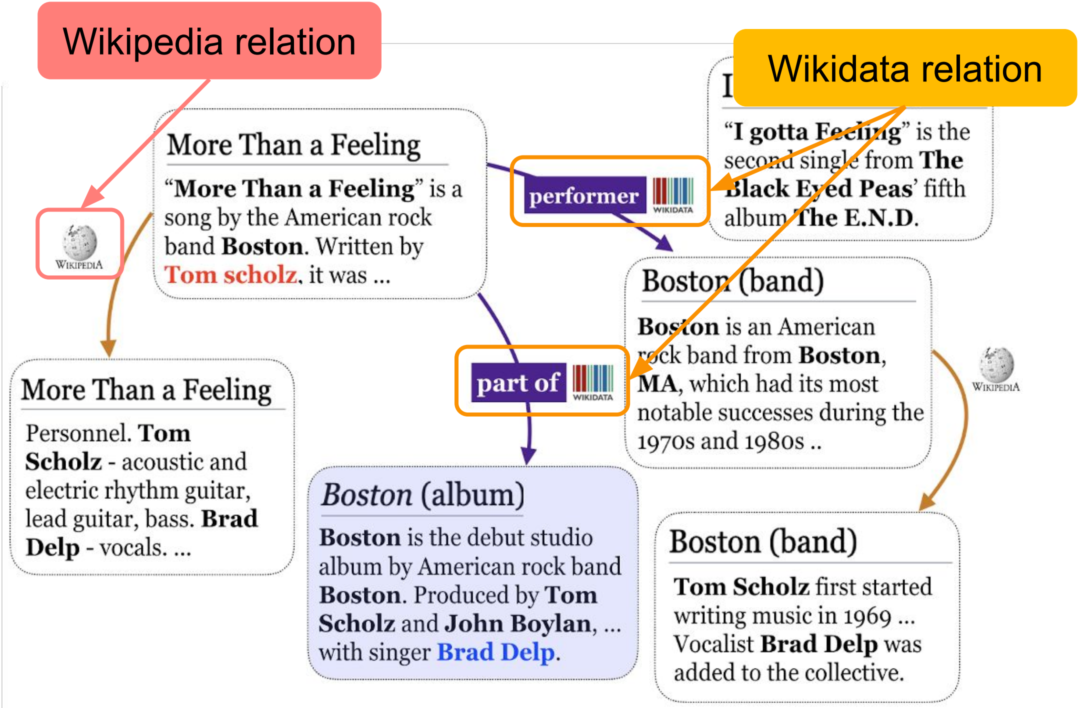
Question: Who sang More than a Feeling by Boston?



Min et al. "Knowledge-guided Text Retrieval and Reading for Open Domain Question Answering." 2019.

Graph-based Methods w/ external knowledge

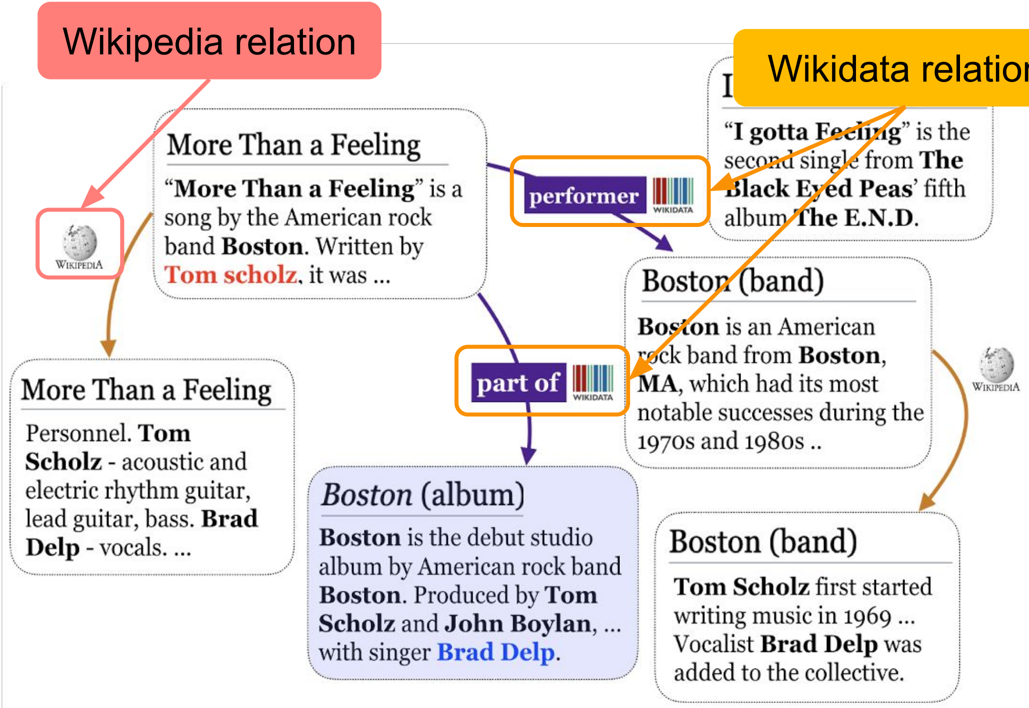
Question: Who sang More than a Feeling by Boston?

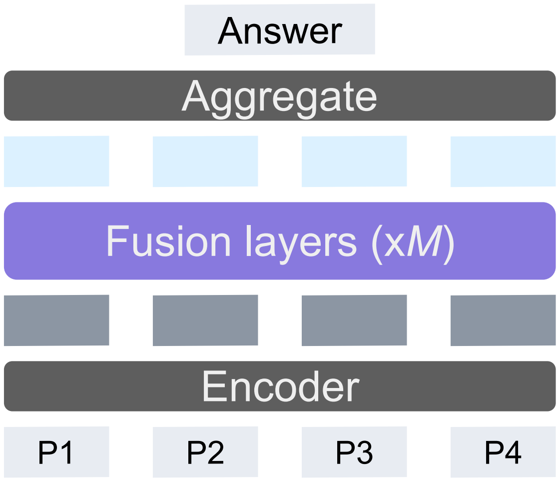


Min et al. "Knowledge-guided Text Retrieval and Reading for Open Domain Question Answering." 2019.

Graph-based Methods w/ external knowledge

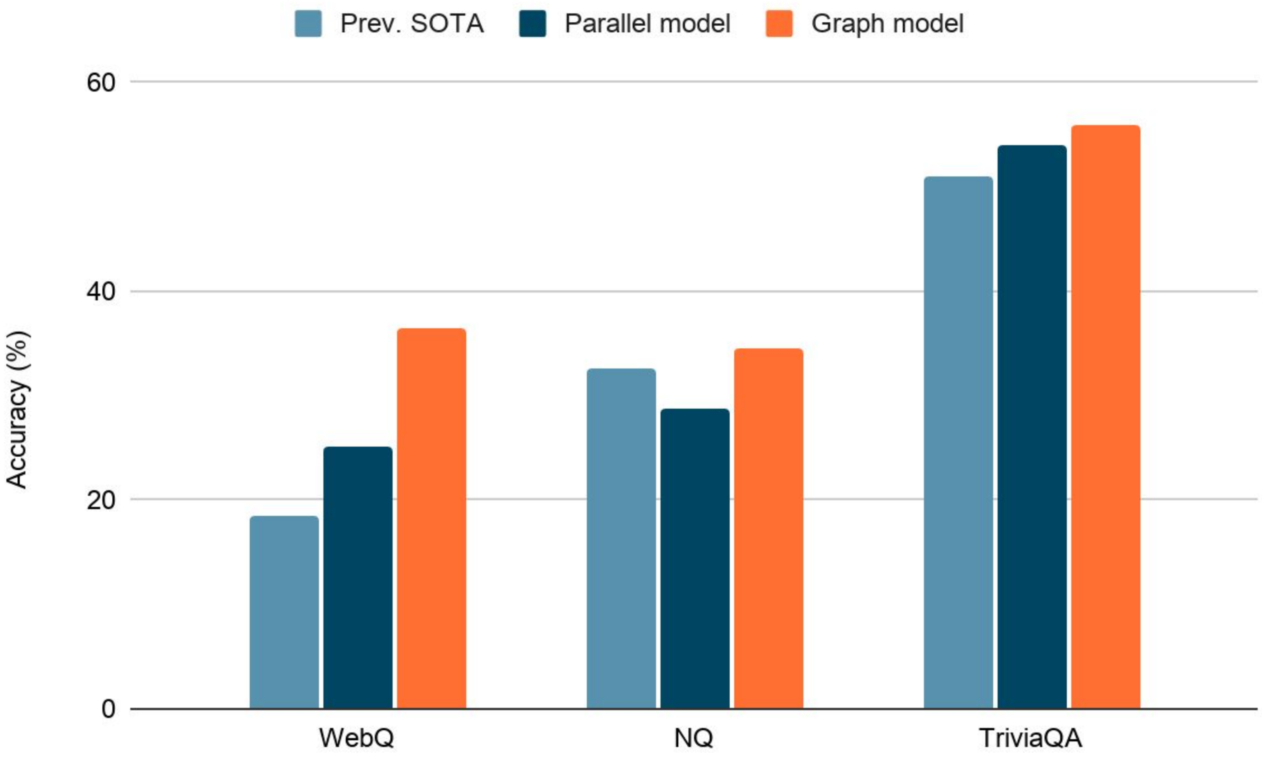
Question: Who sang More than a Feeling by Boston?





Min et al. "Knowledge-guided Text Retrieval and Reading for Open Domain Question Answering." 2019.

Graph-based Methods w/ external knowledge



Min et al. "Knowledge-guided Text Retrieval and Reading for Open Domain Question Answering." 2019.

Summary

- **Hierarchical modeling** leverages the natural hierarchy of the long document
- **Graph-based methods** uses a graph propagation to update the representation of chunks over a chain of chunks in the document
- Graph-based methods can also leverage **external knowledge**