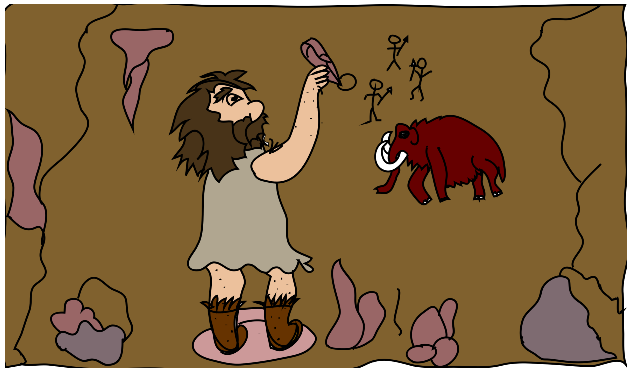
Entity Skeletons for Visual Storytelling

Ruo-Ping*, Khyathi Chandu*, Alan W Black

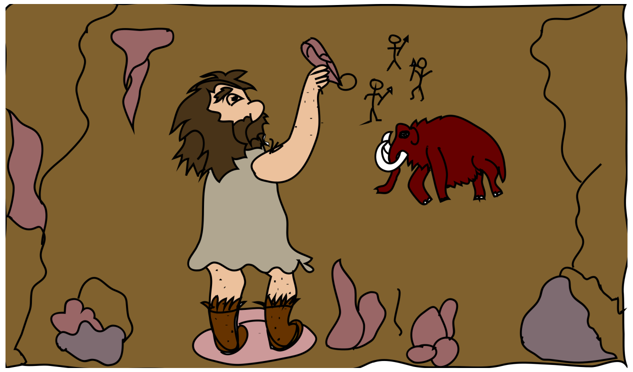
Overview

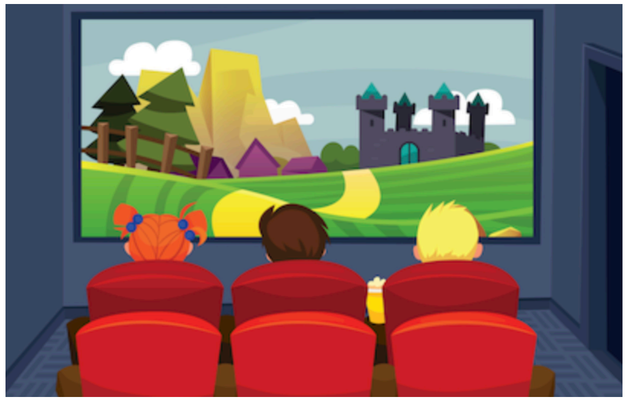
- **Content as a Narrative Property**
- **Task Definition**
- **Dataset**
- **Models**
  - Anchor Extraction
  - Anchor Informed Generation
- **Results**
  - Qualitative and Quantitative
  - Human Evaluation

History of Narratives



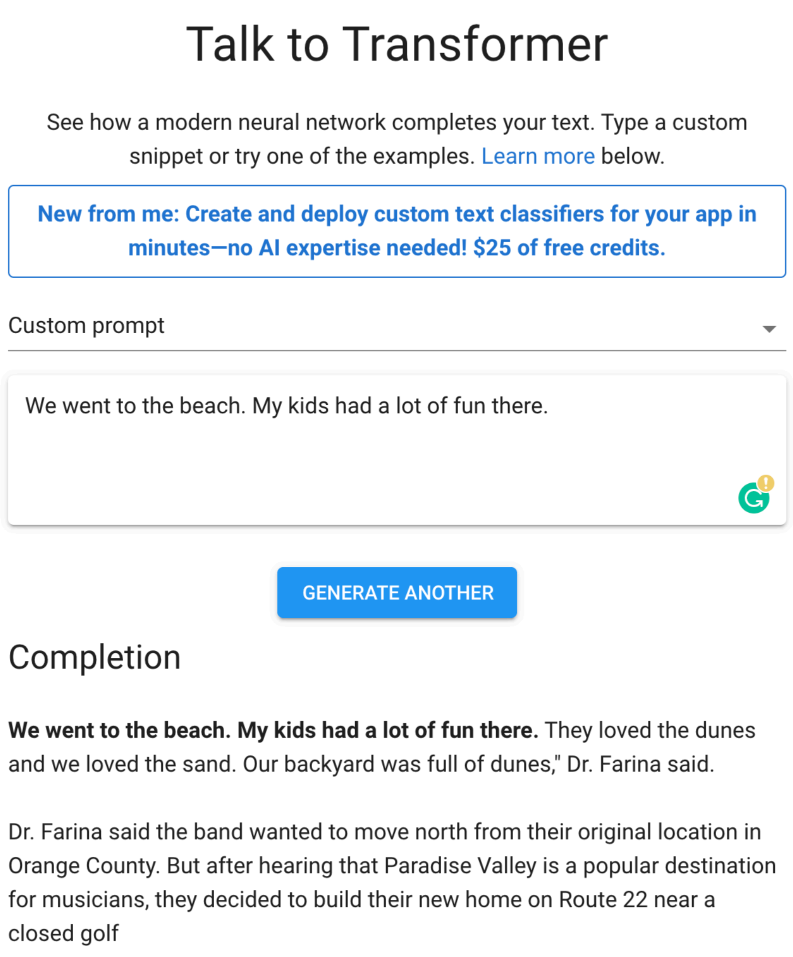
History of Narratives





@magpiessoftserve
vegan corn almond:

Recent Advancements



See how a modern neural network completes your text. Type a custom snippet or try one of the examples. Learn more below.

New from me: Create and deploy custom text classifiers for your app in minutes—no AI expertise needed! $25 of free credits.

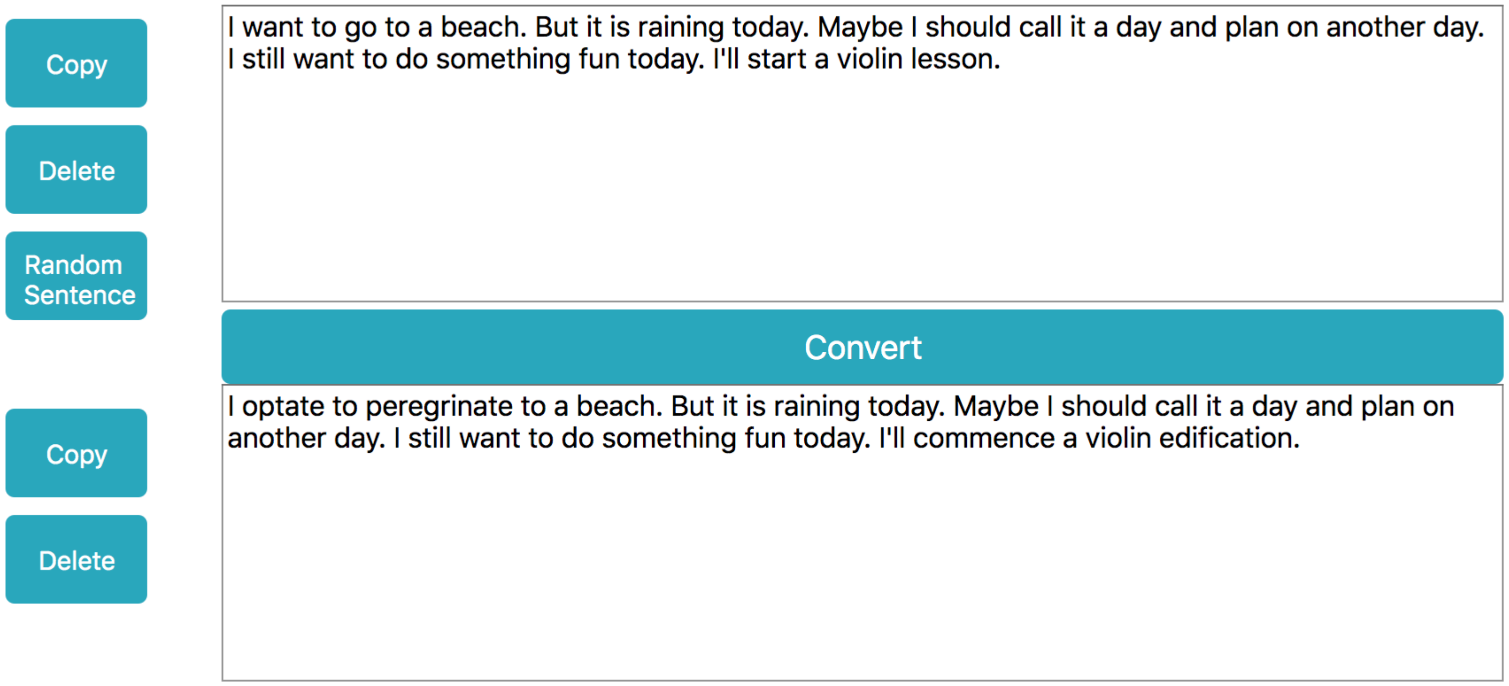
Custom prompt

We went to the beach. My kids had a lot of fun there.

Completion

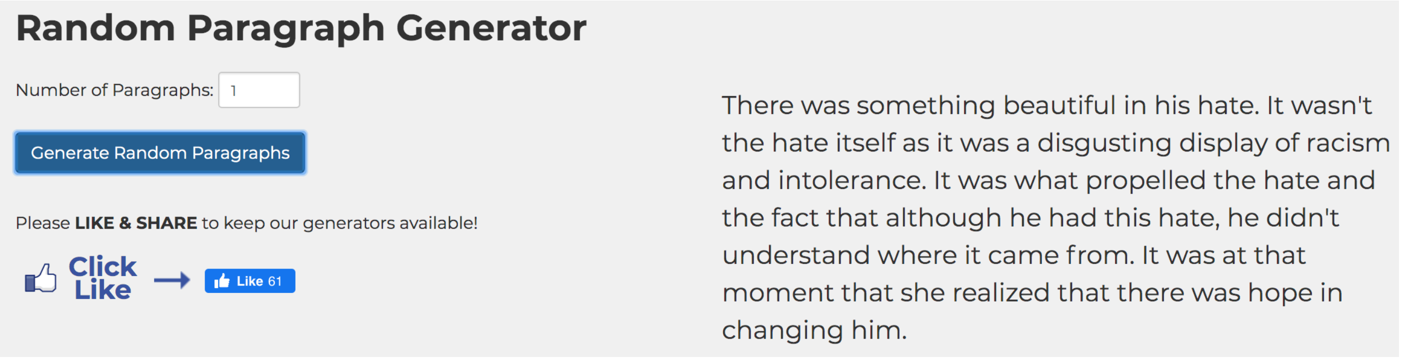
We went to the beach. My kids had a lot of fun there. They loved the dunes and we loved the sand. Our backyard was full of dunes," Dr. Farina said.

Dr. Farina said the band wanted to move north from their original location in Orange County. But after hearing that Paradise Valley is a popular destination for musicians, they decided to build their new home on Route 22 near a closed golf



I want to go to a beach. But it is raining today. Maybe I should call it a day and plan on another day. I still want to do something fun today. I'll start a violin lesson.

I optate to peregrinate to a beach. But it is raining today. Maybe I should call it a day and plan on another day. I still want to do something fun today. I'll commence a violin edification.



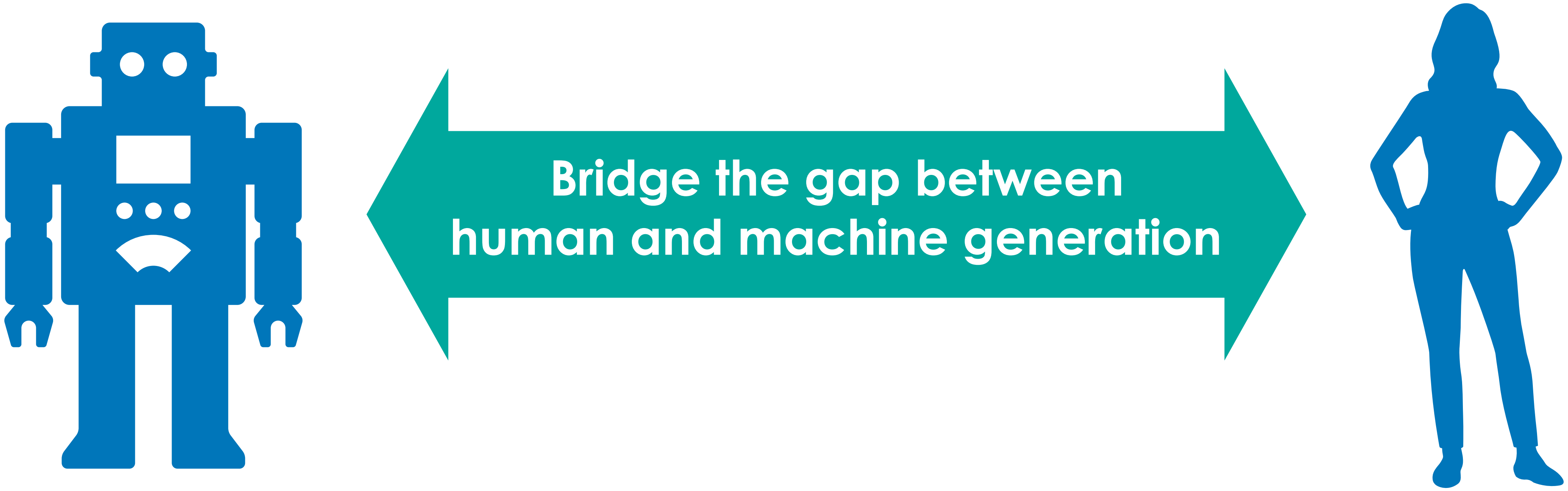
Random Paragraph Generator
Number of Paragraphs: 1
Generate Random Paragraphs
Please LIKE & SHARE to keep our generators available!
Click Like
Like 61

There was something beautiful in his hate. It wasn't the hate itself as it was a disgusting display of racism and intolerance. It was what propelled the hate and the fact that although he had this hate, he didn't understand where it came from. It was at that moment that she realized that there was hope in changing him.

https://talktotransformer.com/

https://randomwordgenerator.com/paragraph.php

https://www.csgenerator.com/



Bridge the gap between human and machine generation

What makes a narrative effective?

Content - Relevance

We went to the beach.
My kids had a lot of fun there.
There were a lot of palm trees.
We stayed in a resort.

We went to the library.
I love reading books.
I borrowed a lot of them.

http://picturebite.com/Views-Sample-Albums

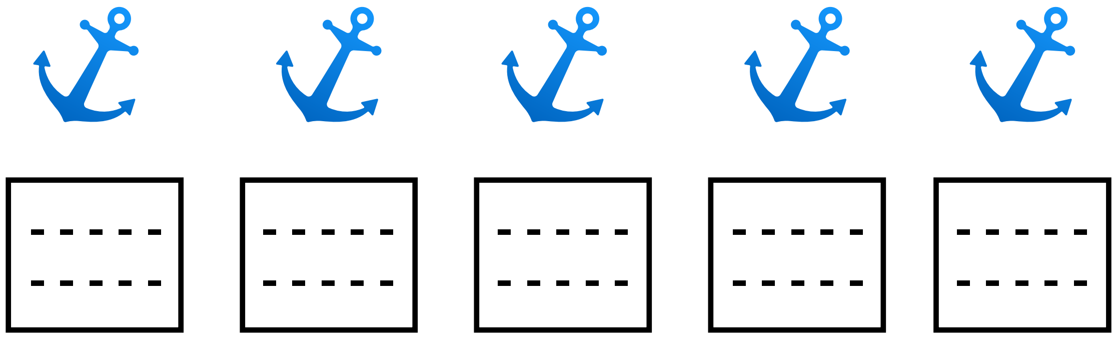
Content - Relevance

Entities
- Beach
- Kids
- Resort

Entities
- Library
- Student
- Books

Anchoring Framework

Fine-grained Entity Skeleton



$$Input : I_i \; and \; E_i = \{e_i^{(1)}, e_i^{(2)}, \ldots, e_i^{(k)}\}$$

$$Output : N_i = \{s_i^{(1)}, s_i^{(2)}, \ldots, s_i^{(k)}\}$$

Provides full guidance to each individual unit of narrative text

Task Definition

- **Task**: Introducing entities in visual stories
- **Data**:

$$\mathsf{S} = \{S_1, \ldots, S_n\}$$

$$S_i = \{(I_i^{(1)}, x_i^{(1)}, y_i^{(1)}), \ldots, (I_i^{(5)}, x_i^{(5)}, y_i^{(5)})\}$$

- **Input**: Sequence of Images, Descriptions in Isolation (DII)
- **Output**: Stories in Sequences (SIS)
- **Anchors**: Entities

Dataset

- Visual Storytelling [1]
- Descriptions in Isolation (DII) absent for 25% of images

	Train	Val	Test
# Stories	40,155	4,990	5,055
#Images	200,775	24,950	25,275
#without DII	40,876	4,973	5,195

[1] Huang, Ting-Hao Kenneth, et al. "Visual storytelling." Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.

Entity Anchors

- Entity Skeleton: defined as a linear chain of entities and referring expressions.
- Coreference chains are extracted from Stanford CoreNLP

Entity Anchors: 3 Forms

- Surface Form Coreference Chains

  $\{c_1, c_2, \ldots, c_5\}$

- Nominalized Coreference Chains

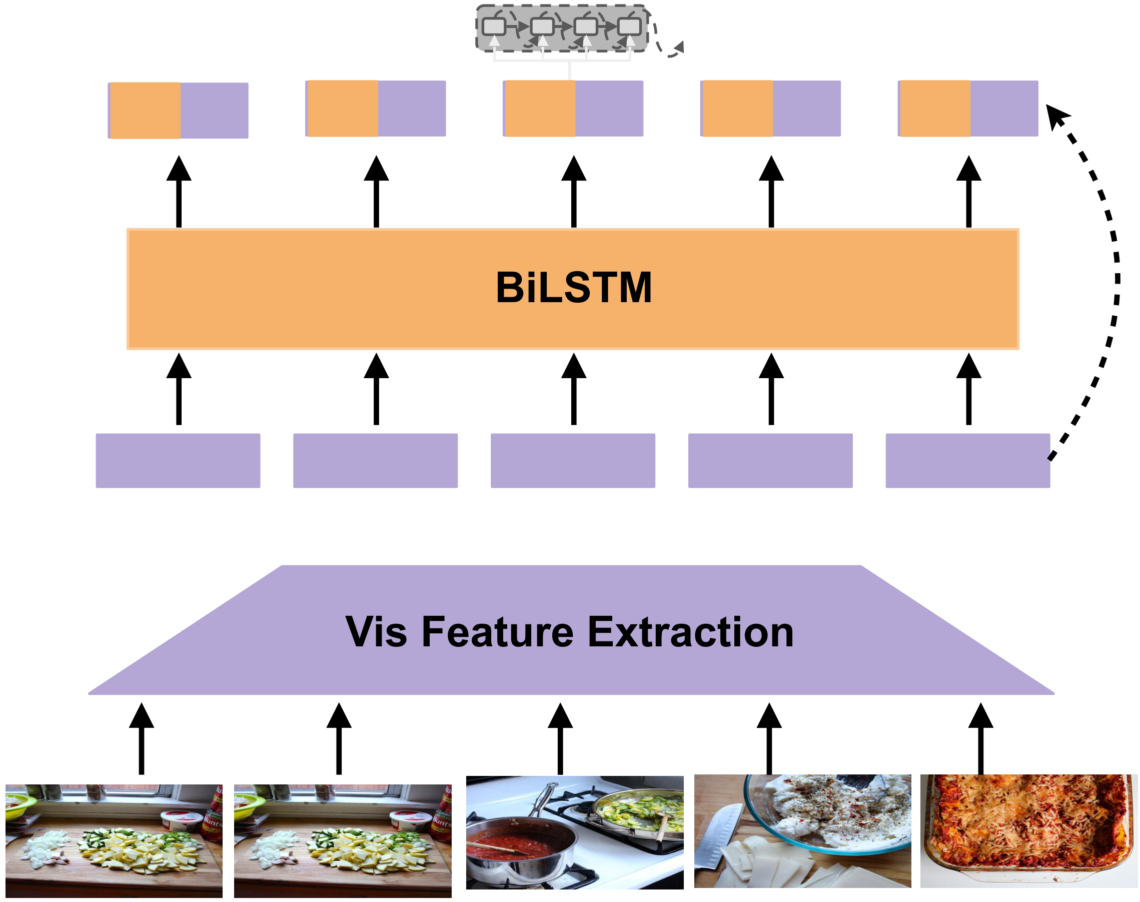
  $\{[p,h]_1, \ldots, [p,h]_5\} \qquad p, h \in \{0,1\}$

- Abstract Coreference Chains

  $\{person, location, misc, object\}$

Anchor Informed Generation

(1) Baseline: Glocal Context Model



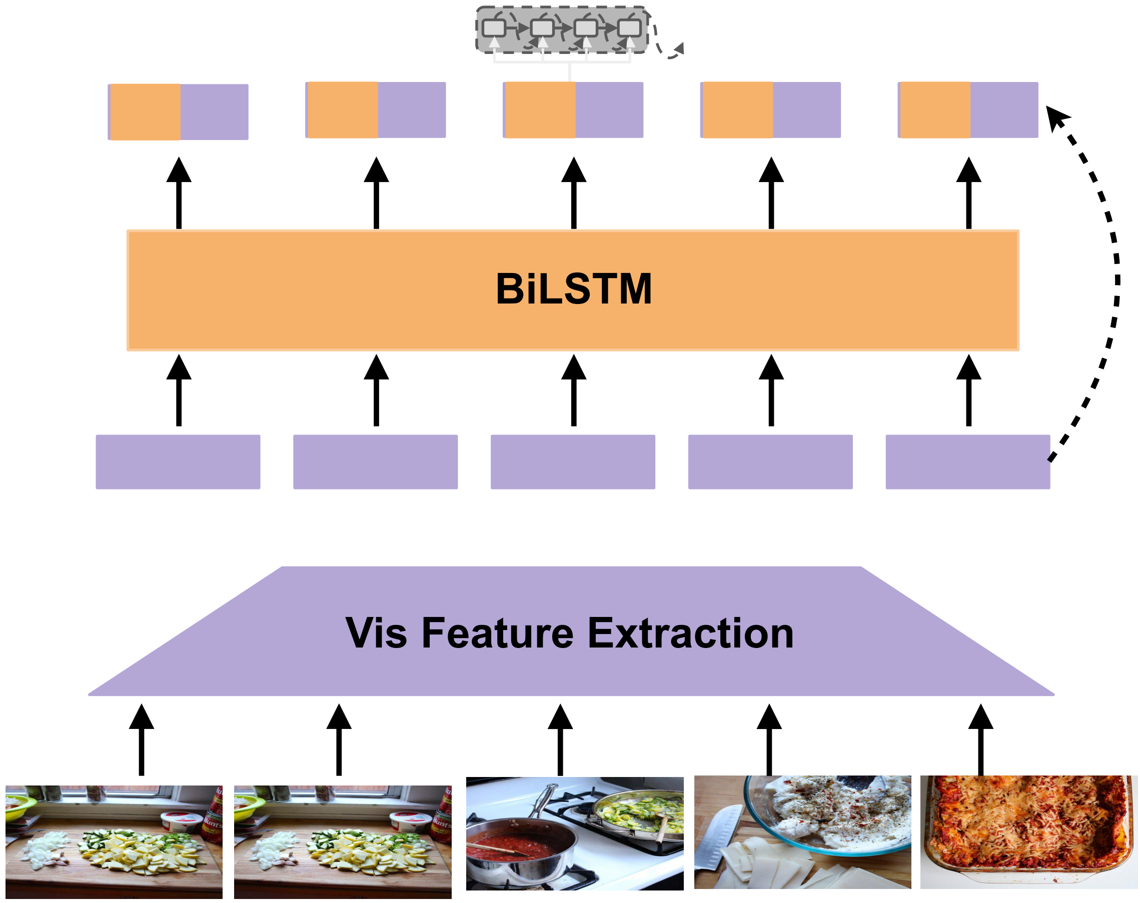
$$\boldsymbol{l}_t = \text{ResNet}(\boldsymbol{I}_t)$$

$$\boldsymbol{g}_t = \text{Bi-LSTM}([l_1, l_2 \ldots l_5]_t)$$

$$\hat{\boldsymbol{w}}_t \sim \prod_{\tau} Pr(\hat{\boldsymbol{w}}_t^{\tau} \,|\, \hat{\boldsymbol{w}}_t^{<\tau}, \boldsymbol{l}_t, \boldsymbol{g}_t)$$

Anchor Informed Generation

(2) Baseline: Skeleton Informed Local Context Model



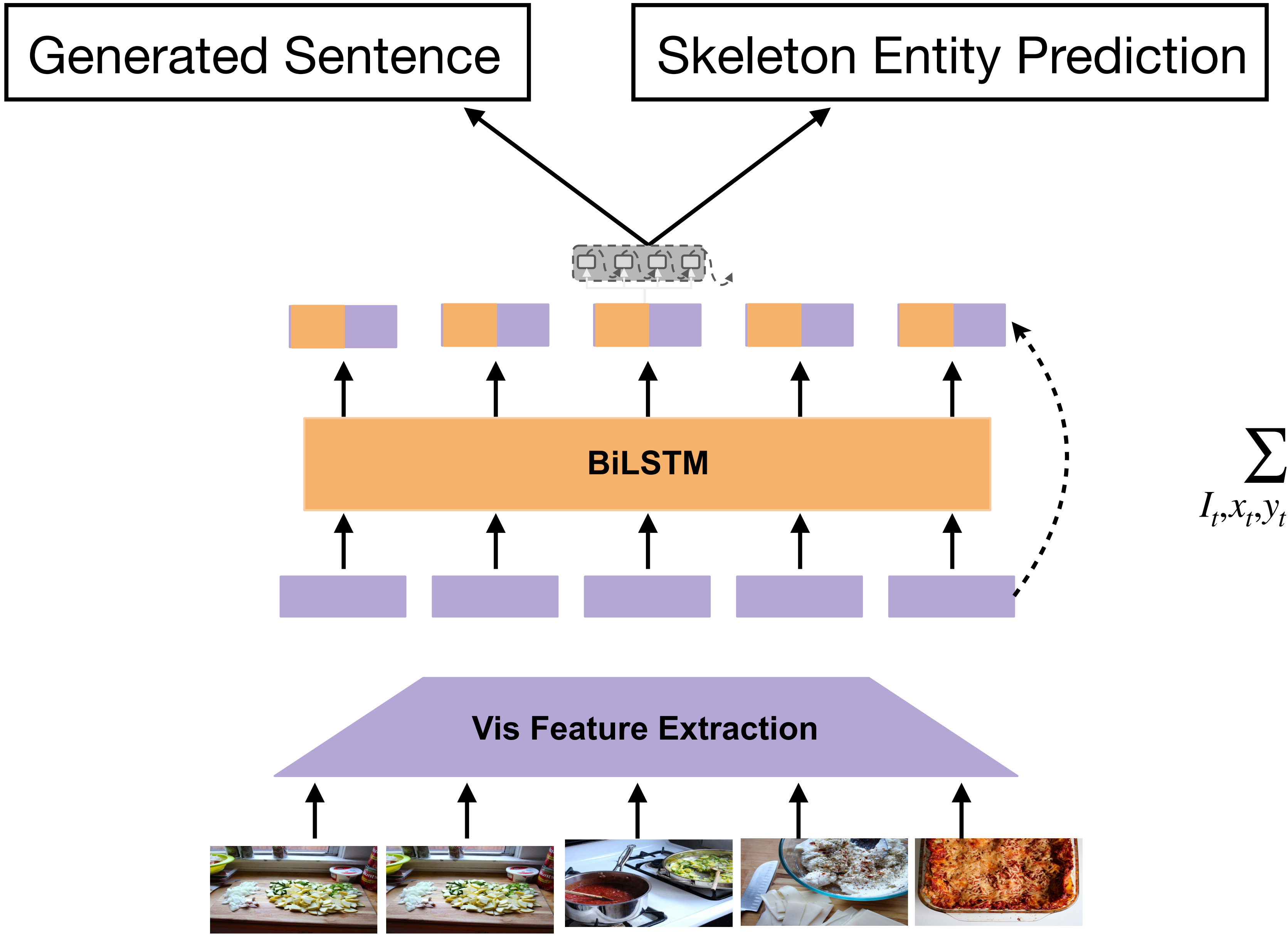
$$\boldsymbol{l}_t = \text{ResNet}(\boldsymbol{I}_t)$$

$$\boldsymbol{g}_t = \text{Bi-LSTM}([l_1, l_2 \ldots l_5]_t)$$

$$\hat{\boldsymbol{w}}_t \sim \prod_{\tau} Pr(\hat{\boldsymbol{w}}_t^{\tau} \mid \hat{\boldsymbol{w}}_t^{<\tau}, \boldsymbol{l}_t, \boldsymbol{g}_t, \boxed{\boldsymbol{k}_t})$$

Anchor Informed Generation

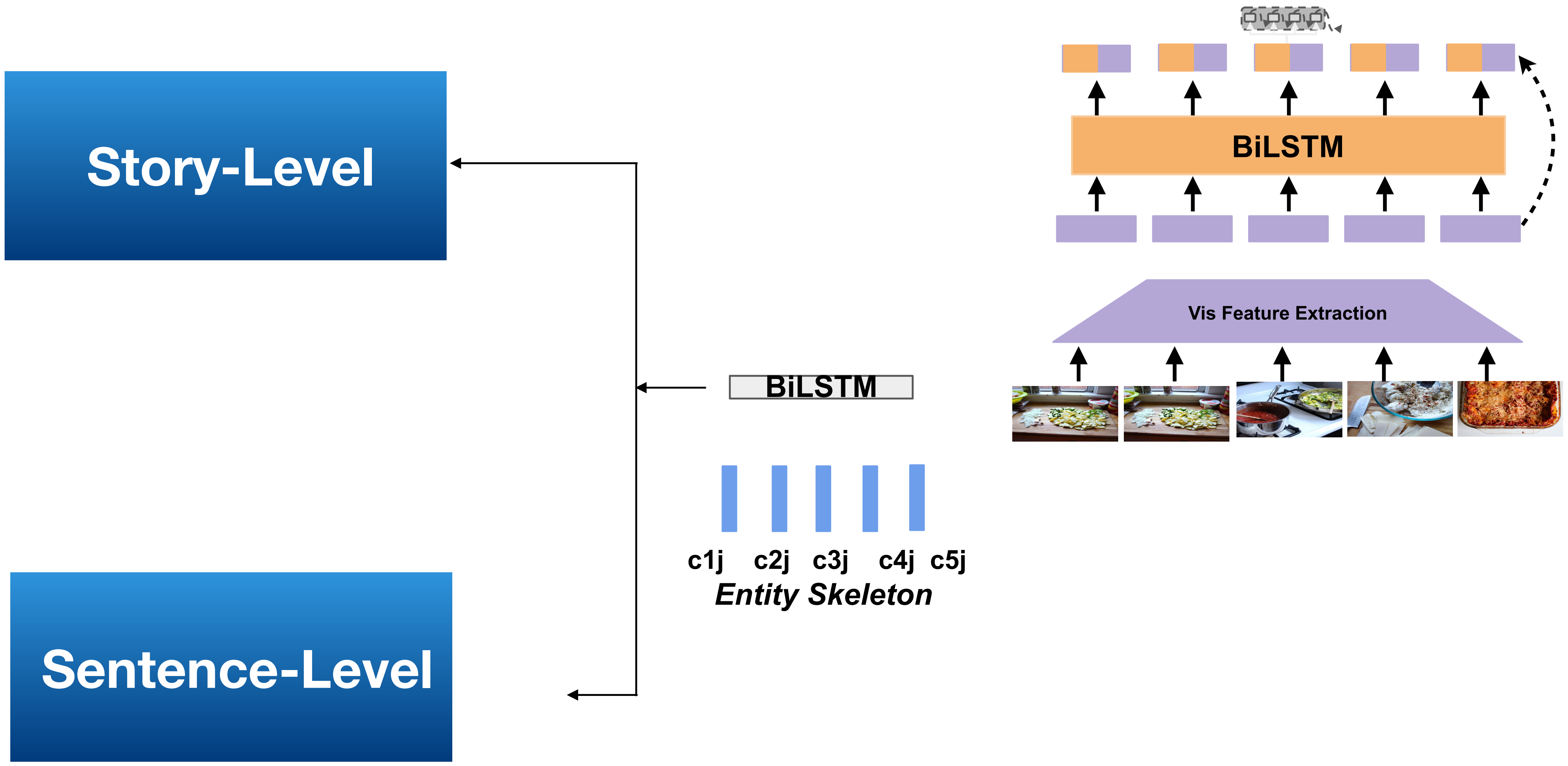
(3) Multitasking Skeleton Prediction



$$\sum_{I_t, x_t, y_t \in S} \alpha L_1(I_t, y_t) + (1 - \alpha) L_2(I_t, y_t, k_t)$$

Anchor Informed Generation

(4) Hierarchical Glocal Model



Story-Level

Sentence-Level

BiLSTM

c1j c2j c3j c4j c5j

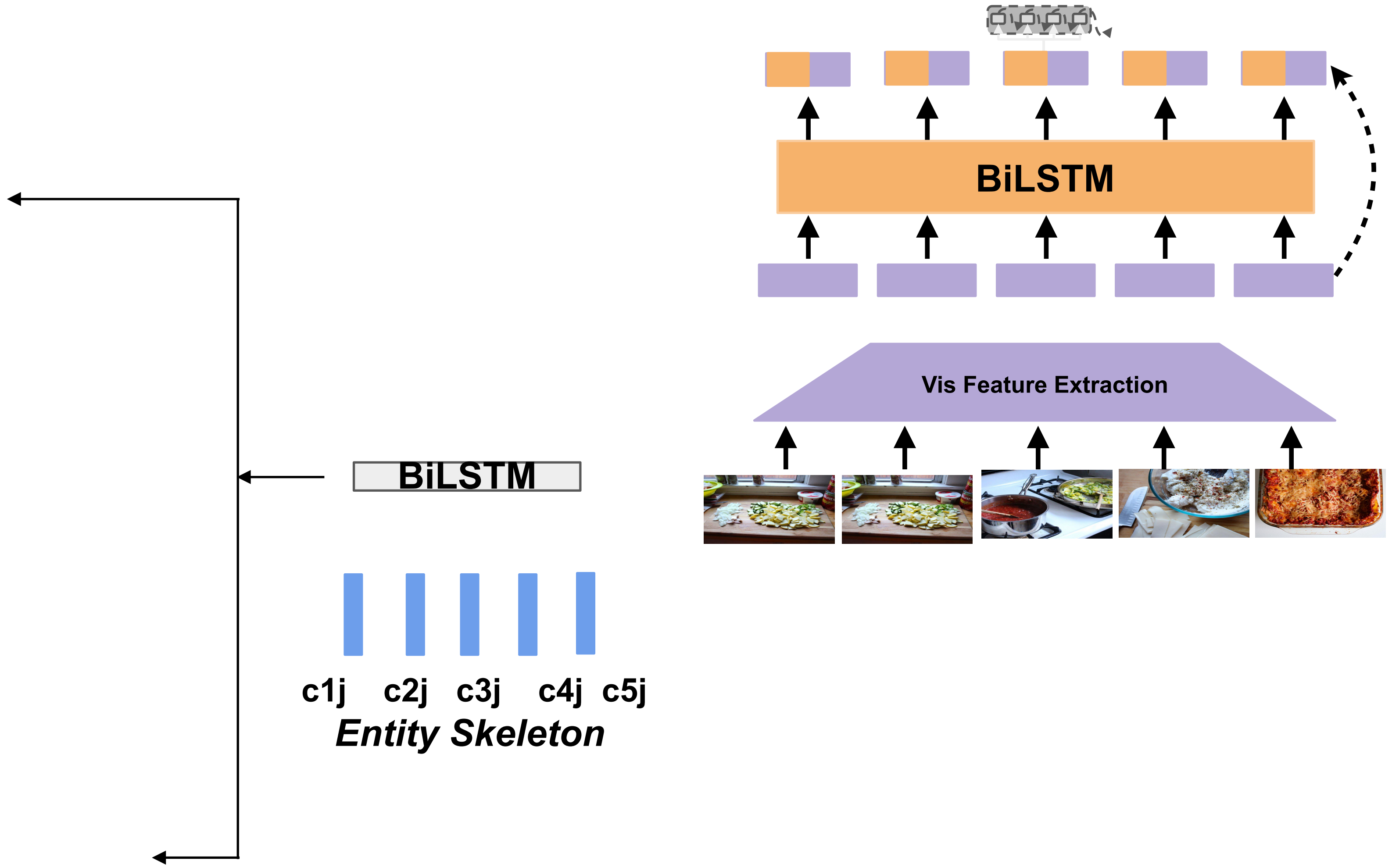
Entity Skeleton

BiLSTM

Vis Feature Extraction

Anchor Informed Generation

(4) Hierarchical Glocal Model



BiLSTM

Vis Feature Extraction

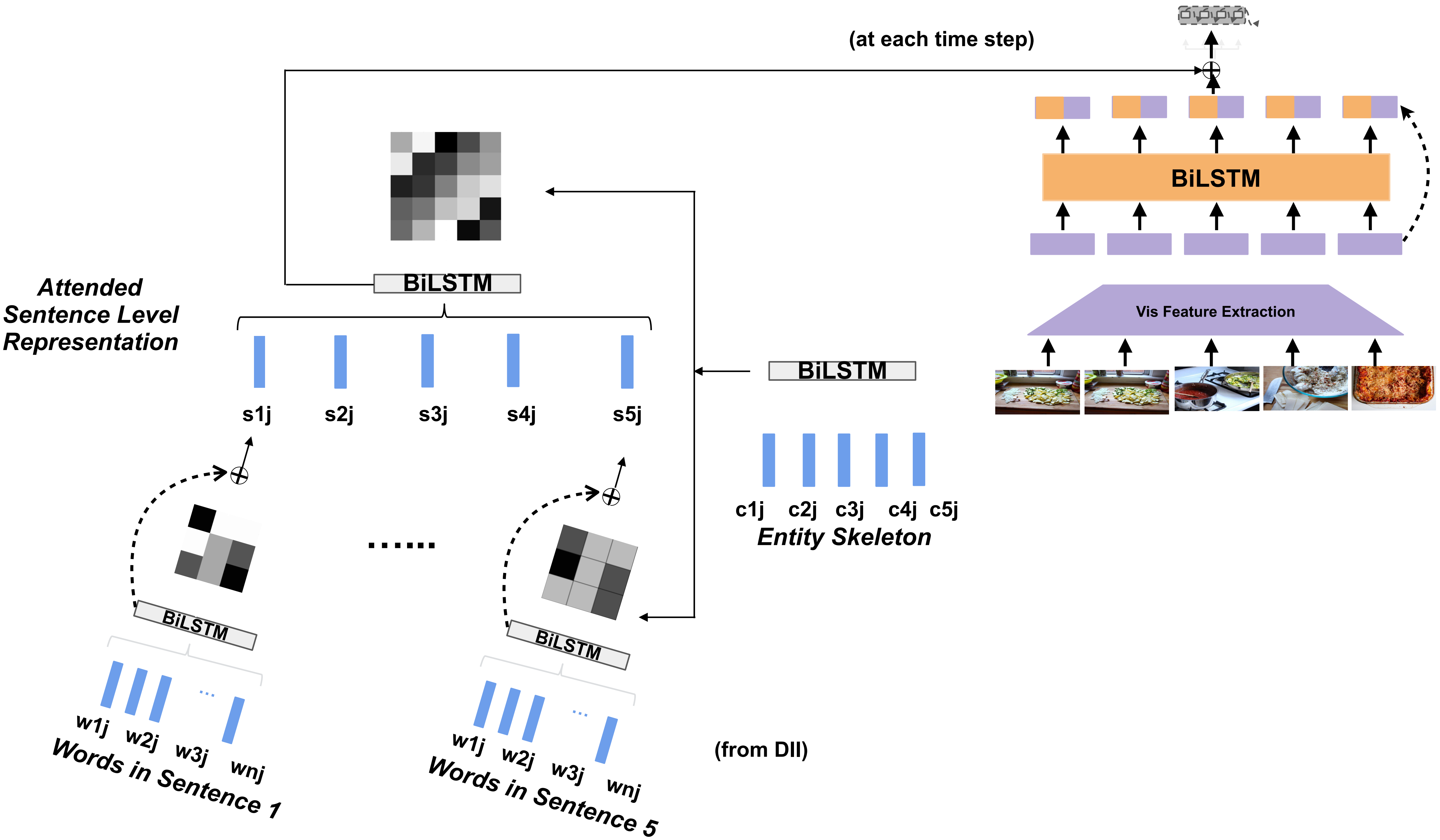
BiLSTM

c1j c2j c3j c4j c5j

Entity Skeleton

Anchor Informed Generation

(4) Hierarchical Glocal Model



Evaluation: Quantitative

Quantitative Experimental Results

Models	Skeleton Form	METEOR	Distance	Avg # of Distinct Entities
Baseline	None	27.93	1.02	0.4971
Baseline + Skeleton	Surface	27.66	1.02	0.5014
MTG (α=0.5)	Surface	27.44	1.02	0.9554
MTG (α=0.4)	Surface	27.59	1.02	1.1013
MTG (α=0.2)	Surface	27.54	1.01	0.9989
MTG (α=0.5)	Nominalization	30.52	1.12	0.5545
MTG (α=0.5)	Abstract	27.67	1.01	0.5115
Glocal Attention	Surface	28.93	1.01	0.8963

Ground Truth: 0.7944

Evaluation: Quantitative

Quantitative Experimental Results

Models	Skeleton Form	METEOR	Distance	Avg # of Distinct Entities
Baseline	None	27.93	1.02	0.4971
Baseline + Skeleton	Surface	27.66	1.02	0.5014
MTG (α=0.5)	Surface	27.44	1.02	0.9554
MTG (α=0.4)	Surface	27.59	1.02	1.1013
MTG (α=0.2)	Surface	27.54	1.01	0.9989
MTG (α=0.5)	Nominalization	30.52	1.12	0.5545
MTG (α=0.5)	Abstract	27.67	1.01	0.5115
Glocal Attention	Surface	28.93	1.01	0.8963

Ground Truth:
0.7944

Evaluation: Quantitative

Quantitative Experimental Results

Models	Skeleton Form	METEOR	Distance	Avg # of Distinct Entities
Baseline	None	27.93	1.02	0.4971
Baseline + Skeleton	Surface	27.66	1.02	0.5014
MTG (α=0.5)	Surface	27.44	1.02	0.9554
MTG (α=0.4)	Surface	27.59	1.02	1.1013
MTG (α=0.2)	Surface	27.54	1.01	0.9989
MTG (α=0.5)	Nominalization	30.52	1.12	0.5545
MTG (α=0.5)	Abstract	27.67	1.01	0.5115
Glocal Attention	Surface	28.93	1.01	0.8963

Ground Truth: 0.7944

Evaluation: Quantitative

Quantitative Experimental Results

Models	Skeleton Form	METEOR	Distance	Avg # of Distinct Entities
Baseline	None	27.93	1.02	0.4971
Baseline + Skeleton	Surface	27.66	1.02	0.5014
MTG (α=0.5)	Surface	27.44	1.02	0.9554
MTG (α=0.4)	Surface	27.59	1.02	1.1013
MTG (α=0.2)	Surface	27.54	1.01	0.9989
MTG (α=0.5)	Nominalization	**30.52**	1.12	0.5545
MTG (α=0.5)	Abstract	27.67	1.01	0.5115
Glocal Attention	Surface	**28.93**	1.01	**0.8963**

Ground Truth: 0.7944

Human Evaluation

- Preference Testing for Hierarchical Glocal Model
  - 82% over Baseline
  - 64% over Multitasking Model

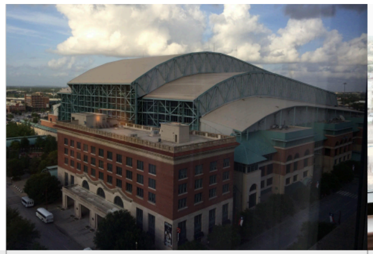
Evaluation: Qualitative

Qualitative Analysis

Models						*Phenomena*
SIS	**we** went to the stadium early to eat and sight see before the game .	the view was incredible . you could see the entire city .	**we** got to our seats , and could n't believe how close to the field they were .	**we** could see all the action .	once the national anthem was sung , and the first pitch was thrown , the excitement began . **it** was a great game !	
Baseline Model	the city was a great place to visit .	i had a great time .	there were many people there .	**we** got to see a lot of cool things .	**it** was a lot of fun .	- Characters in the story are mentioned as “many people” instead of “we” (sentence 3).
Glocal Hierarchical Attention Model	**we** saw the **building** was packed .	i was excited to see my favorite team .	**we** were all excited to see the **game** .	**we** all got together to watch .	**it** was a great **game** .	+ characters (‘we’ and ‘it’) were introduced at the right time + Important entities were mentioned (building, game)

Evaluation: Qualitative

Qualitative Analysis

Models						*Phenomena*
SIS	**we** went to the stadium early to eat and sight see before the game .	the view was incredible . you could see the entire city .	**we** got to our seats , and could n't believe how close to the field they were .	**we** could see all the action .	once the national anthem was sung , and the first pitch was thrown , the excitement began . **it** was a great game !	
Baseline Model	the city was a great place to visit .	i had a great time .	there were many people there .	**we** got to see a lot of cool things .	**it** was a lot of fun .	- Characters in the story are mentioned as “many people” instead of “we” (sentence 3).
Glocal Hierarchical Attention Model	**we** saw the **building** was packed .	i was excited to see my favorite team .	**we** were all excited to see the **game** .	**we** all got together to watch .	**it** was a great **game** .	+ characters (‘we’ and ‘it’) were introduced at the right time + Important entities were mentioned (building, game)

Takeaways

- Improves Relevance component of visual storytelling
- Improves Controllability in generation
- Step towards interpretability with respect to intermediate representation

Thank You

Contact: kchandu@cs.cmu.edu