Automatic Paraphrase Collection and Identification in Twitter

Wuwei Lan, Siyu Qiu, Hua He, Wei Xu

The Ohio State University

Penn University of Pennsylvania

University of Maryland

What is paraphrase?

What is paraphrase?

Willy Wonka was famous for his delicious candy.
Children and adults loved to eat it.

What is paraphrase?

Willy Wonka was famous for his delicious candy.
Children and adults loved to eat it.

Willy Wonka was known throughout the world because people enjoyed eating the tasty candy he made.

What is paraphrase?

Willy Wonka was famous for his delicious candy.
Children and adults loved to eat it.

Willy Wonka was known throughout the world because people enjoyed eating the tasty candy he made.

What is paraphrase?

Willy Wonka was famous for his delicious candy.
Children and adults loved to eat it.

Willy Wonka was known throughout the world because people
enjoyed eating the tasty candy he made.

What is paraphrase?

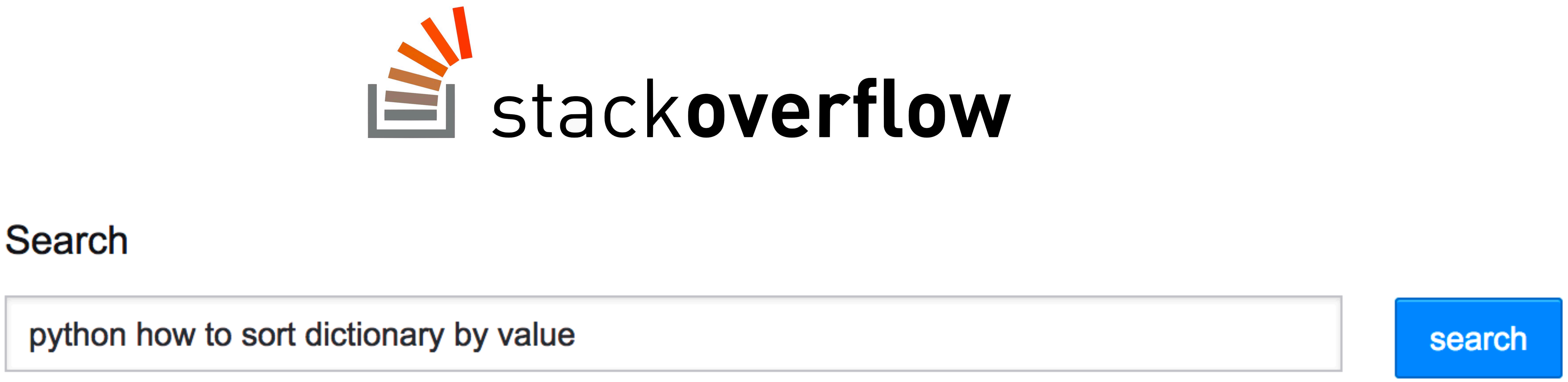
Willy Wonka was famous for his delicious candy.
Children and adults loved to eat it.

Willy Wonka was known throughout the world because people enjoyed eating the tasty candy he made.

Paraphrase Application: duplicate question identification

Paraphrase Application: duplicate question identification

Paraphrase Application: duplicate question identification



Paraphrase Application: duplicate question identification

Paraphrase Application: duplicate question identification

stackoverflow

Search

python how to sort dictionary by value

search

Q: Sort a Python dictionary by value

Q: How to sort a Python dictionary by value?

Q: Python how to sort a dictionary by value in reverse order

Paraphrase

Paraphrase Application: duplicate question identification

stackoverflow

Search

python how to sort dictionary by value

search

2477 votes | 38 answers | Q: Sort a Python dictionary by value

12 votes | 2 answers | Q: How to sort a Python dictionary by value?

-4 votes | 3 answers | Q: Python how to sort a dictionary by value in reverse order

Paraphrase

Paraphrase Application: duplicate question identification

stackoverflow

Search

python how to sort dictionary by value

search

2477 votes | 38 answers

Q: Sort a Python dictionary by value

, but **how** can I **sort** based on the values? Note: I have read Stack Overflow question **How** do I **sort** a list of dictionaries by values of the **dictionary** in **Python**? and probably could change my code to have … I have a **dictionary** of values read from two fields in a database: a string field and a numeric field. The string field is unique, so that is the key of the **dictionary**. I can **sort** on the keys …

python sorting dictionary

asked Mar 5 '09 by Gern Blanston

12 votes | 2 answers

Q: How to sort a Python dictionary by value?

-4 votes | 3 answers

Q: Python how to sort a dictionary by value in reverse order

Paraphrase

Paraphrase Application: question answering

Paraphrase Application: question answering

IBM Watson

Paraphrase Application: question answering

IBM Watson

[Question]

In May 1898 Portugal celebrated the 400th anniversary of this explorer's arrival in India

Paraphrase Application: question answering

IBM Watson

[Question]

In May 1898 Portugal celebrated the 400th anniversary of this explorer's arrival in India

[Supporting Evidence]

On the 27th of May 1498, Vasco da Gama landed in Kappad Beach

Paraphrase Application: question answering

IBM Watson

[Question]

In May 1898 Portugal celebrated the 400th anniversary of this explorer’s arrival in India

celebrated

May 1898 | 400th anniversary | Portugal

arrival in

India

explorer

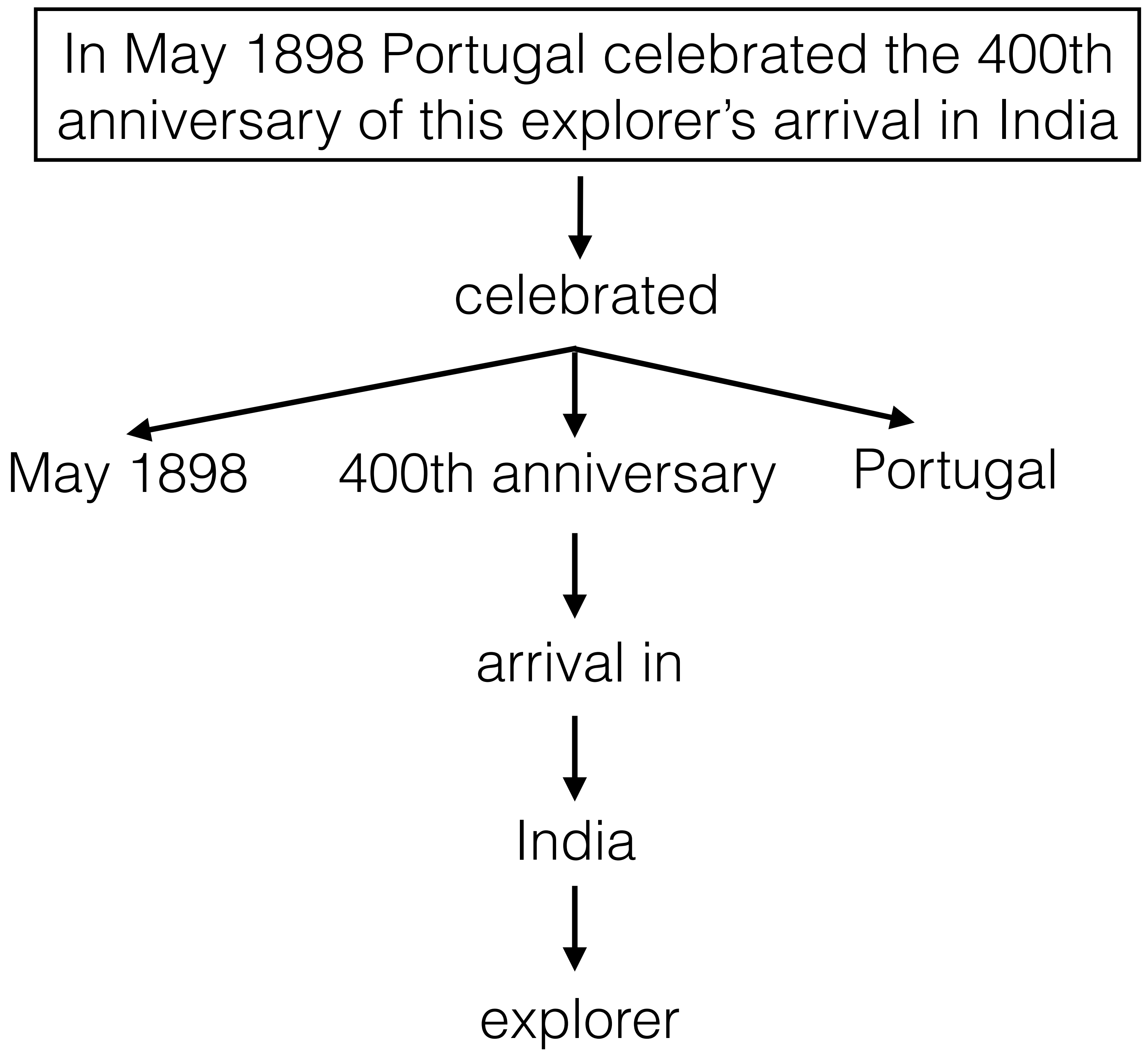
[Supporting Evidence]

On the 27th of May 1498, Vasco da Gama landed in Kappad Beach

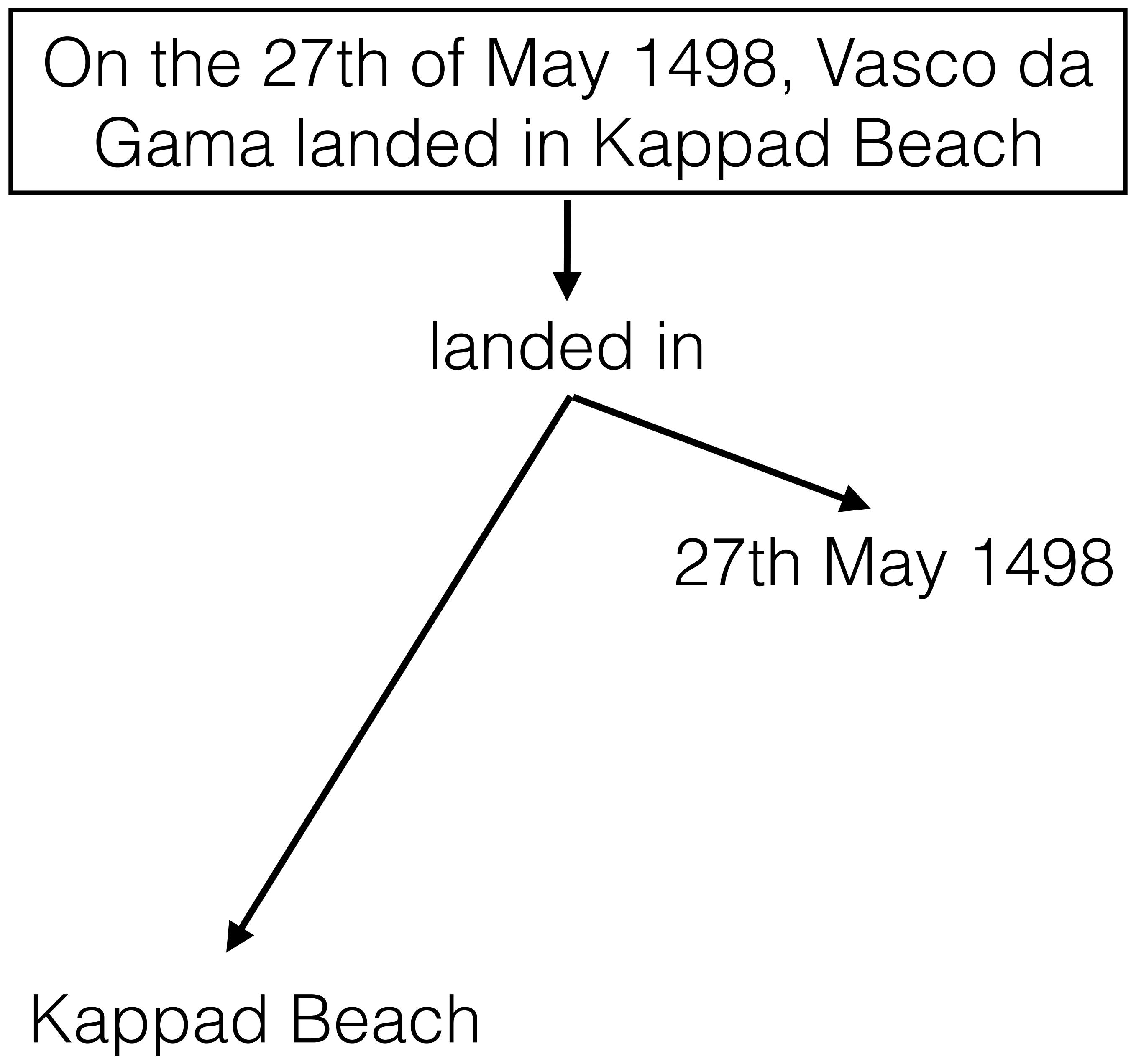
Paraphrase Application: question answering

IBM Watson

[Question]



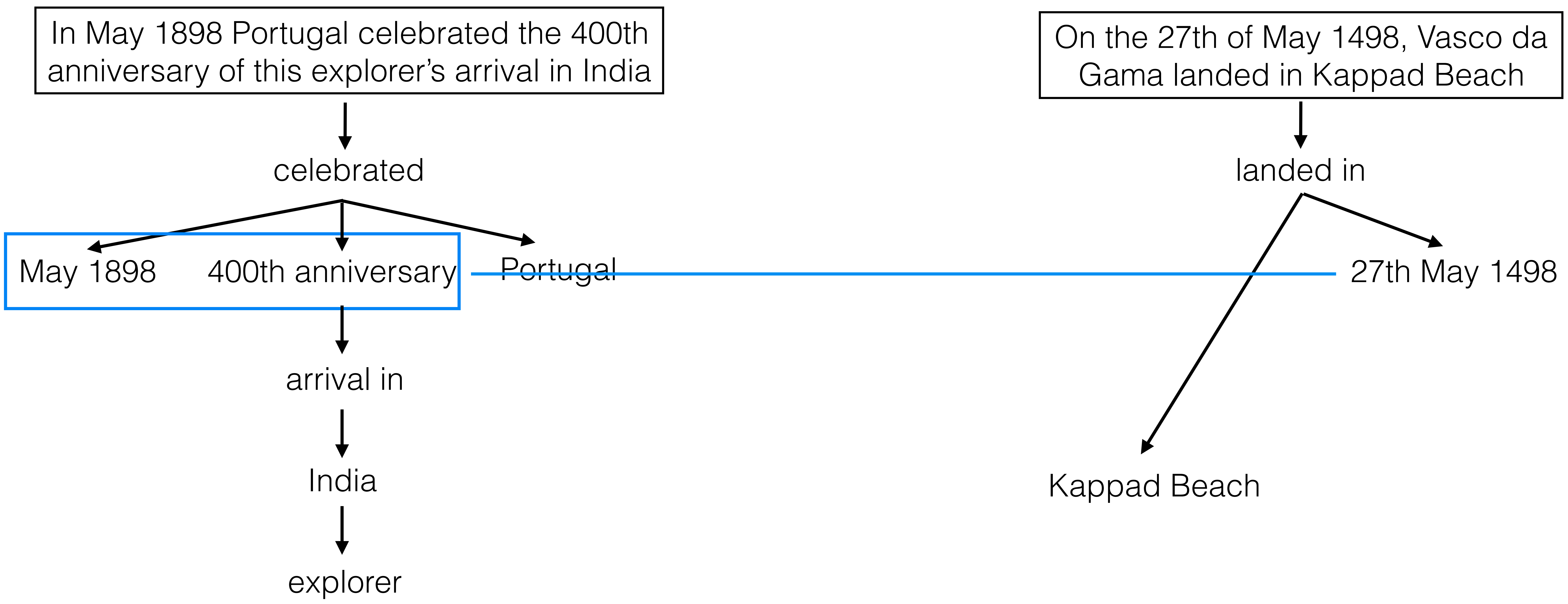
[Supporting Evidence]



Paraphrase Application: question answering

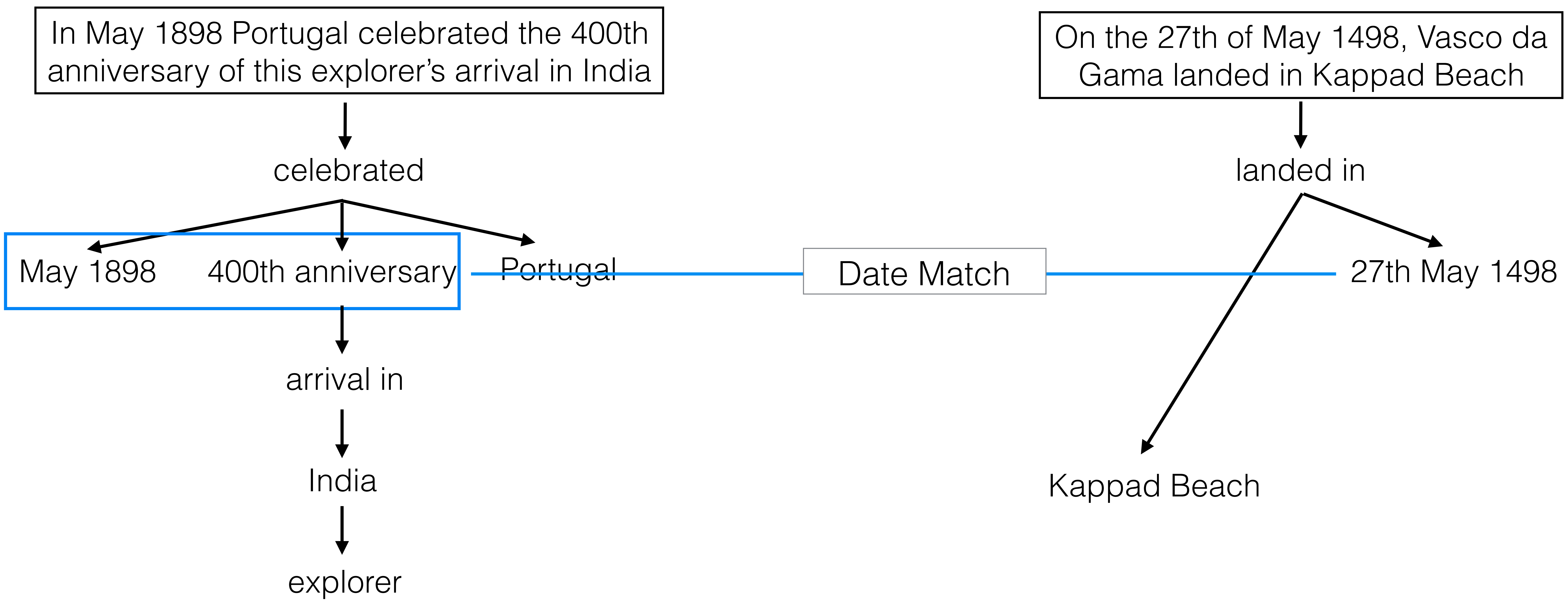
IBM Watson

[Question] **[Supporting Evidence]**



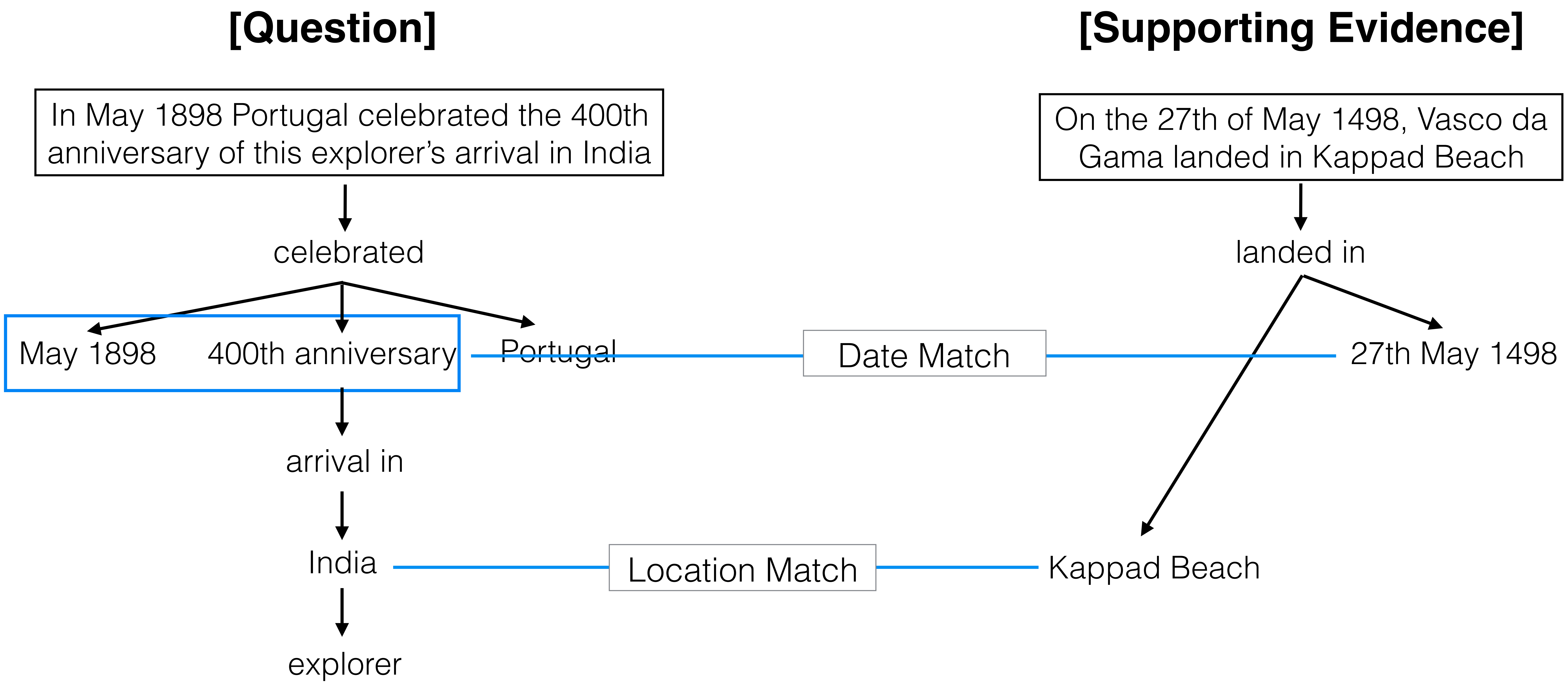
Paraphrase Application: question answering

[Question] **[Supporting Evidence]**



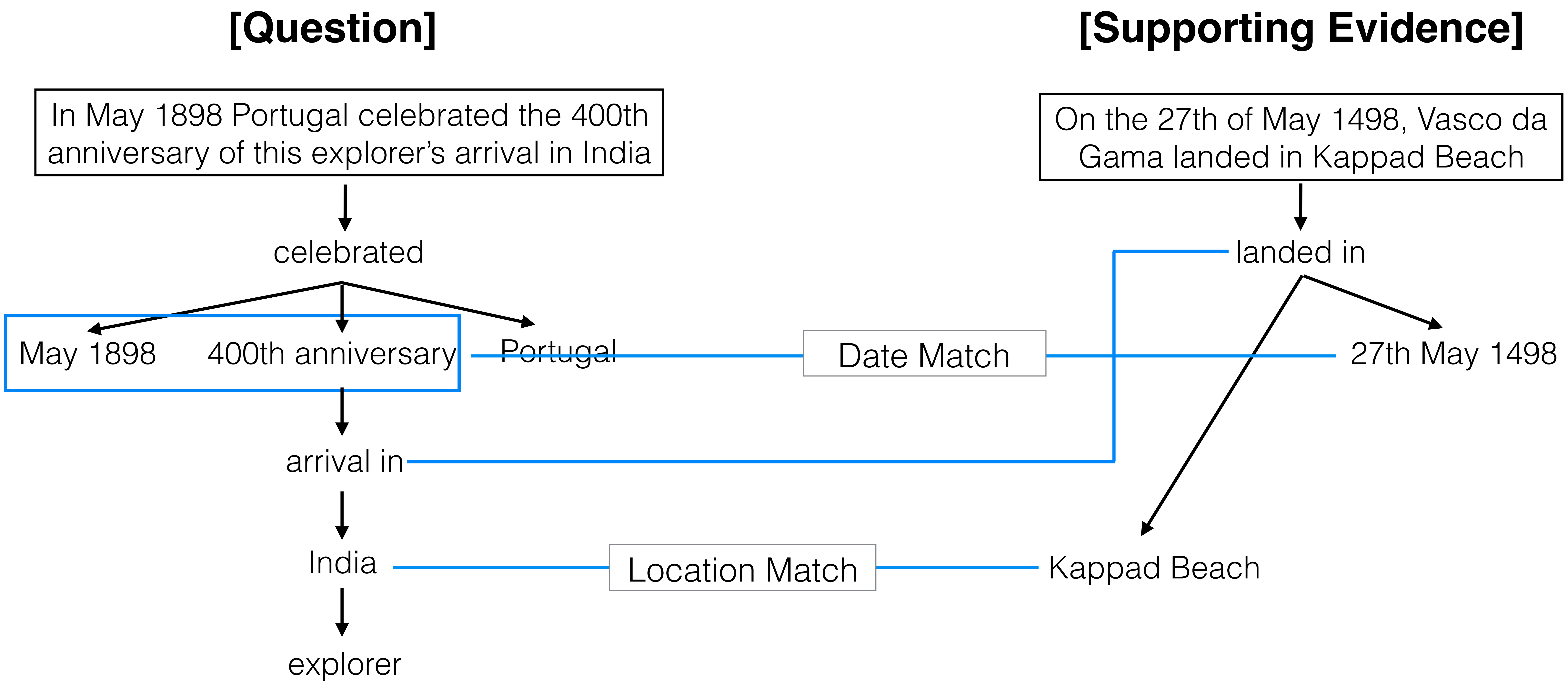
Paraphrase Application: question answering

IBM Watson



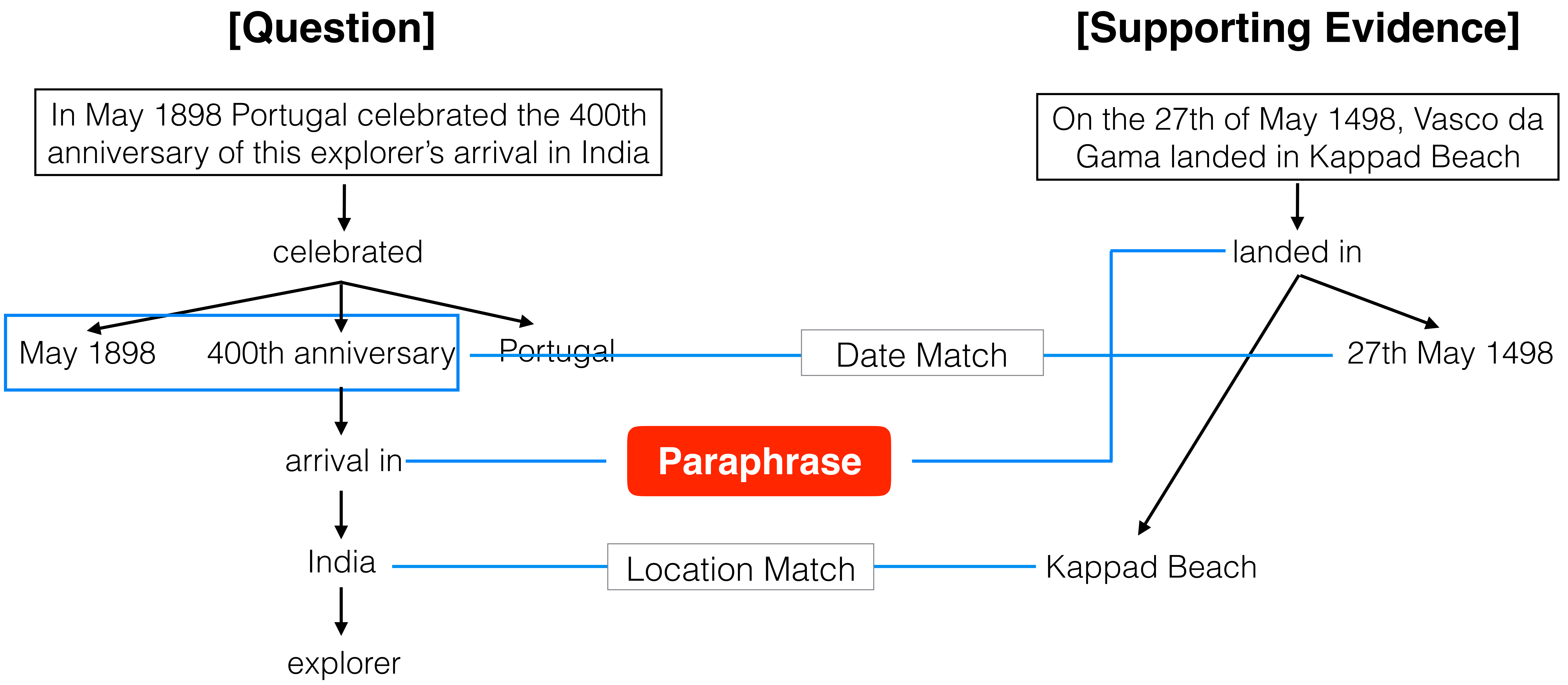
Paraphrase Application: question answering

IBM Watson



Paraphrase Application: question answering

IBM Watson



Paraphrases?

https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html

BOB DYLAN
A NOBEL AWARD
IN LITERATURE

'CAPITAL OF KITSCH'
APPRECIATING ABSURDITY
AND ANTIQUITY IN SKOPJE

'IT'S JUST WORDS'
THAT'S WHAT TRUMP SAYS.
THESE 2 WOMEN DIFFER.

The New York Times

Trump gets it right about one thing

Nicholas Kristof

ISIS turning drones into airborne explosives

U.S. must now find ways to thwart small aircraft not only spying but killing

Mourning a beloved king

Filipinos back tough approach

Despite scorn abroad, polls show support for Duterte's harsh measures

Paraphrases?

https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html


BOB DYLAN A NOBEL AWARD IN LITERATURE
'CAPITAL OF KITSCH' APPRECIATING ABSURDITY AND ANTIQUITY IN SKOPJE
'IT'S JUST WORDS' THAT'S WHAT TRUMP SAYS. THESE 2 WOMEN DIFFER.
The New York Times
Trump gets it right about one thing
Nicholas Kristof
ISIS turning drones into airborne explosives
Filipinos back tough approach

Paraphrases?

https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html


BOB DYLAN A NOBEL AWARD IN LITERATURE
'CAPITAL OF KITSCH' APPRECIATING ABSURDITY AND ANTIQUITY IN SKOPJE
'IT'S JUST WORDS' THAT'S WHAT TRUMP SAYS. THESE 2 WOMEN DIFFER.
The New York Times
Trump gets it right about one thing
Nicholas Kristof
ISIS turning drones into airborne explosives
Filipinos back tough approach

Paraphrases?

https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html


BOB DYLAN A NOBEL AWARD IN LITERATURE
'CAPITAL OF KITSCH' APPRECIATING ABSURDITY AND ANTIQUITY IN SKOPJE
'IT'S JUST WORDS' THAT'S WHAT TRUMP SAYS. THESE 2 WOMEN DIFFER.

The New York Times

Trump gets it right about one thing

Nicholas Kristof

ISIS turning drones into airborne explosives

Filipinos back tough approach

The New York Times @nytimes · 12 Oct 2016
Worries over the health of King Bhumibol Adulyadej are shaking Thailand nyti.ms/2dRzPcr
5 261 144

Career Synchronicity @careersync_now · 12 Oct 2016
Fears for King's Health Shake Thailand ift.tt/2d7frGd

Paraphrases?

https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html


BOB DYLAN A NOBEL AWARD IN LITERATURE

'CAPITAL OF KITSCH' APPRECIATING ABSURDITY AND ANTIQUITY IN SKOPJE

'IT'S JUST WORDS' THAT'S WHAT TRUMP SAYS. THESE 2 WOMEN DIFFER.

The New York Times

Trump gets it right about one thing

Nicholas Kristof

ISIS turning drones into airborne explosives

Filipinos back tough approach

The New York Times @nytimes · 12 Oct 2016

Worries over the health of King Bhumibol Adulyadej are shaking Thailand nyti.ms/2dRzPcr

5 261 144

Career Synchronicity @careersync_now · 12 Oct 2016

Fears for King's Health Shake Thailand ift.tt/2d7frGd

Paraphrase

Paraphrases?

https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html


BOB DYLAN A NOBEL AWARD IN LITERATURE
'CAPITAL OF KITSCH' APPRECIATING ABSURDITY AND ANTIQUITY IN SKOPJE
'IT'S JUST WORDS' THAT'S WHAT TRUMP SAYS. THESE 2 WOMEN DIFFER.

The New York Times

Trump gets it right about one thing

Nicholas Kristof

ISIS turning drones into airborne explosives

Filipinos back tough approach

The New York Times @nytimes · 12 Oct 2016

Worries over the health of King Bhumibol Adulyadej are shaking Thailand

nyti.ms/2dRzPcr

5 261 144

Career Synchronicity @careersync_now · 12 Oct 2016

Fears for King's Health Shake Thailand ift.tt/2d7frGd

Paraphrase

Paraphrases?

https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html


The New York Times

Trump gets it right about one thing

ISIS turning drones into airborne explosives

Filipinos back tough approach

The New York Times @nytimes · 12 Oct 2016
Worries over the health of King Bhumibol Adulyadej are shaking Thailand nyti.ms/2dRzPcr
5 261 144

Career Synchronicity @careersync_now · 12 Oct 2016
Fears for King's Health Shake Thailand ift.tt/2d7frGd

Paraphrase

Herbert Buchsbaum @herbertnyt · 12 Oct 2016
New bulletin from Thai palace: King is still on a ventilator and in unstable condition. nyti.ms/2dW1A37

Paraphrases?

https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html


BOB DYLAN A NOBEL AWARD IN LITERATURE

'CAPITAL OF KITSCH' APPRECIATING ABSURDITY AND ANTIQUITY IN SKOPJE

'IT'S JUST WORDS' THAT'S WHAT TRUMP SAYS. THESE 2 WOMEN DIFFER.

The New York Times

INTERNATIONAL EDITION

Trump gets it right about one thing

Nicholas Kristof

ISIS turning drones into airborne explosives

Filipinos back tough approach

The New York Times @nytimes · 12 Oct 2016

Worries over the health of King Bhumibol Adulyadej are shaking Thailand nyti.ms/2dRzPcr

5 261 144

Career Synchronicity @careersync_now · 12 Oct 2016

Fears for King's Health Shake Thailand ift.tt/2d7frGd

Paraphrase

Herbert Buchsbaum @herbertnyt · 12 Oct 2016

New bulletin from Thai palace: King is still on a ventilator and in unstable condition. nyti.ms/2dW1A37

Non-Paraphrase

Paraphrases? We can get many in Twitter

https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html

The New York Times @nytimes · 12 Oct 2016
Worries over the health of King Bhumibol Adulyadej are shaking Thailand nyti.ms/2dRzPcr
5 261 144

Career Synchronicity @careersync_now · 12 Oct 2016
Fears for King's Health Shake Thailand ift.tt/2d7frGd

Herbert Buchsbaum @herbertnyt · 12 Oct 2016
New bulletin from Thai palace: King is still on a ventilator and in unstable condition. nyti.ms/2dW1A37

Paraphrases? We can get many in Twitter

https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html

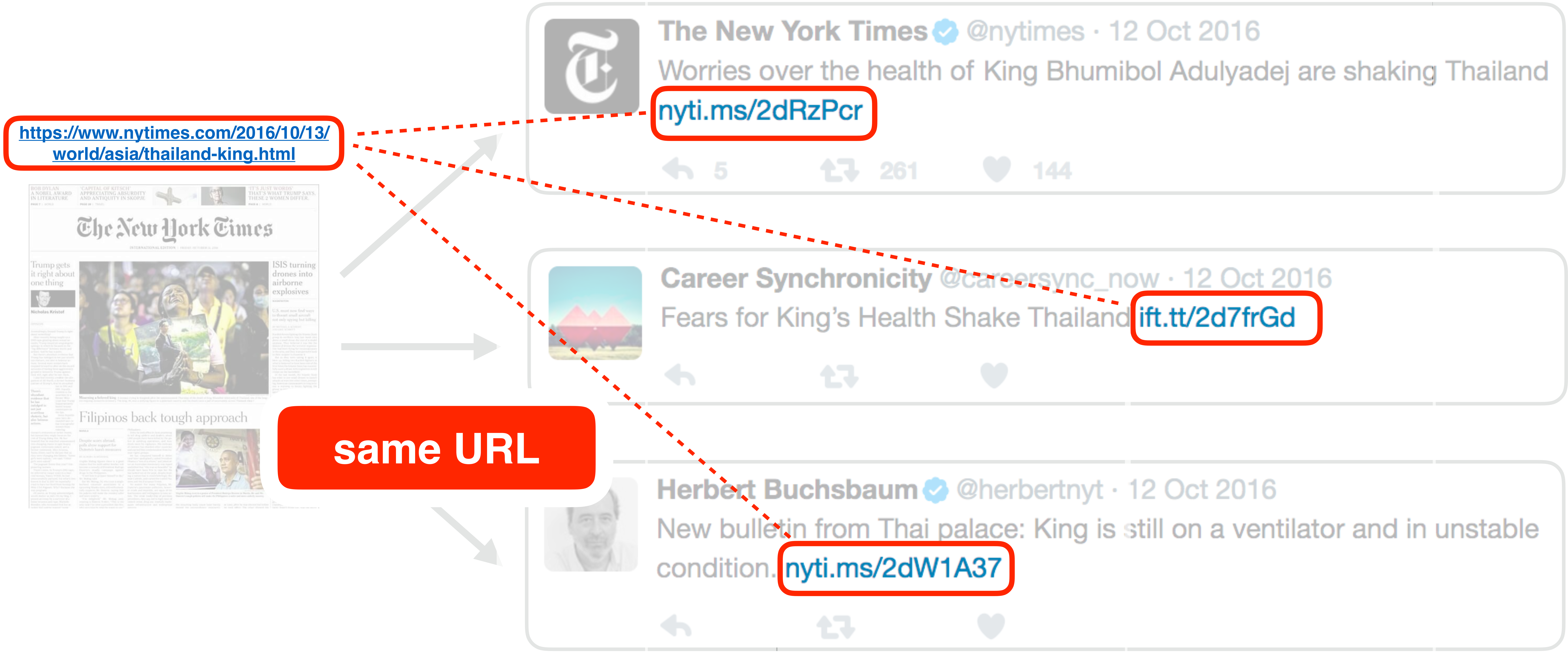

The New York Times
Trump gets it right about one thing
ISIS turning drones into airborne explosives
Filipinos back tough approach

The New York Times @nytimes · 12 Oct 2016
Worries over the health of King Bhumibol Adulyadej are shaking Thailand nyti.ms/2dRzPcr
5 261 144

Career Synchronicity @careersync_now · 12 Oct 2016
Fears for King's Health Shake Thailand ift.tt/2d7frGd

Herbert Buchsbaum @herbertnyt · 12 Oct 2016
New bulletin from Thai palace: King is still on a ventilator and in unstable condition. nyti.ms/2dW1A37

Paraphrases? We can get many in Twitter



Only exist two sentential paraphrase corpora
(which contain meaningful non-paraphrases)

[MSRP[1]]

clustered
news articles

5,801 annotated pairs

[PIT-2015[2]]

Twitter
trending topics

14,035 annotated pairs

[1] Dolan et al., 2004
[2] Xu et al., 2014

Only exist two sentential paraphrase corpora
(which contain meaningful non-paraphrases)

Key for success:

[MSRP[1]]

clustered news articles

5,801 annotated pairs

[PIT-2015[2]]

Twitter trending topics

14,035 annotated pairs

[1] Dolan et al., 2004
[2] Xu et al., 2014

Only exist two sentential paraphrase corpora
(which contain meaningful non-paraphrases)

Key for success:

- narrow the search space

[MSRP[1]]

clustered
news articles

5,801 annotated pairs

[PIT-2015[2]]

Twitter
trending topics

14,035 annotated pairs

[1] Dolan et al., 2004
[2] Xu et al., 2014

Only exist two sentential paraphrase corpora
(which contain meaningful non-paraphrases)

Key for success:

- narrow the search space
- ensure diversity among sentences

[MSRP[1]]

clustered
news articles

5,801 annotated pairs

[PIT-2015[2]]

Twitter
trending topics

14,035 annotated pairs

[1] Dolan et al., 2004
[2] Xu et al., 2014

Only exist two sentential paraphrase corpora
(which contain meaningful non-paraphrases)

Key for success:

- narrow the search space
- ensure diversity among sentences

Also Pitfalls …

[MSRP[1]]

clustered news articles

5,801 annotated pairs

[PIT-2015[2]]

Twitter trending topics

14,035 annotated pairs

[1] Dolan et al., 2004
[2] Xu et al., 2014

Only exist two sentential paraphrase corpora
(which contain meaningful non-paraphrases)

Key for success:

- narrow the search space
- ensure diversity among sentences

Also Pitfalls …

[MSRP[1]]

clustered news articles

5,801 annotated pairs

needed a SVM classifier to select sentences before data annotation

[PIT-2015[2]]

Twitter trending topics

14,035 annotated pairs

[1] Dolan et al., 2004
[2] Xu et al., 2014

Only exist two sentential paraphrase corpora
(which contain meaningful non-paraphrases)

Key for success:

- narrow the search space
- ensure diversity among sentences

Also Pitfalls …

[MSRP[1]]

clustered news articles

5,801 annotated pairs

needed a SVM classifier to select sentences before data annotation

[PIT-2015[2]]

Twitter trending topics

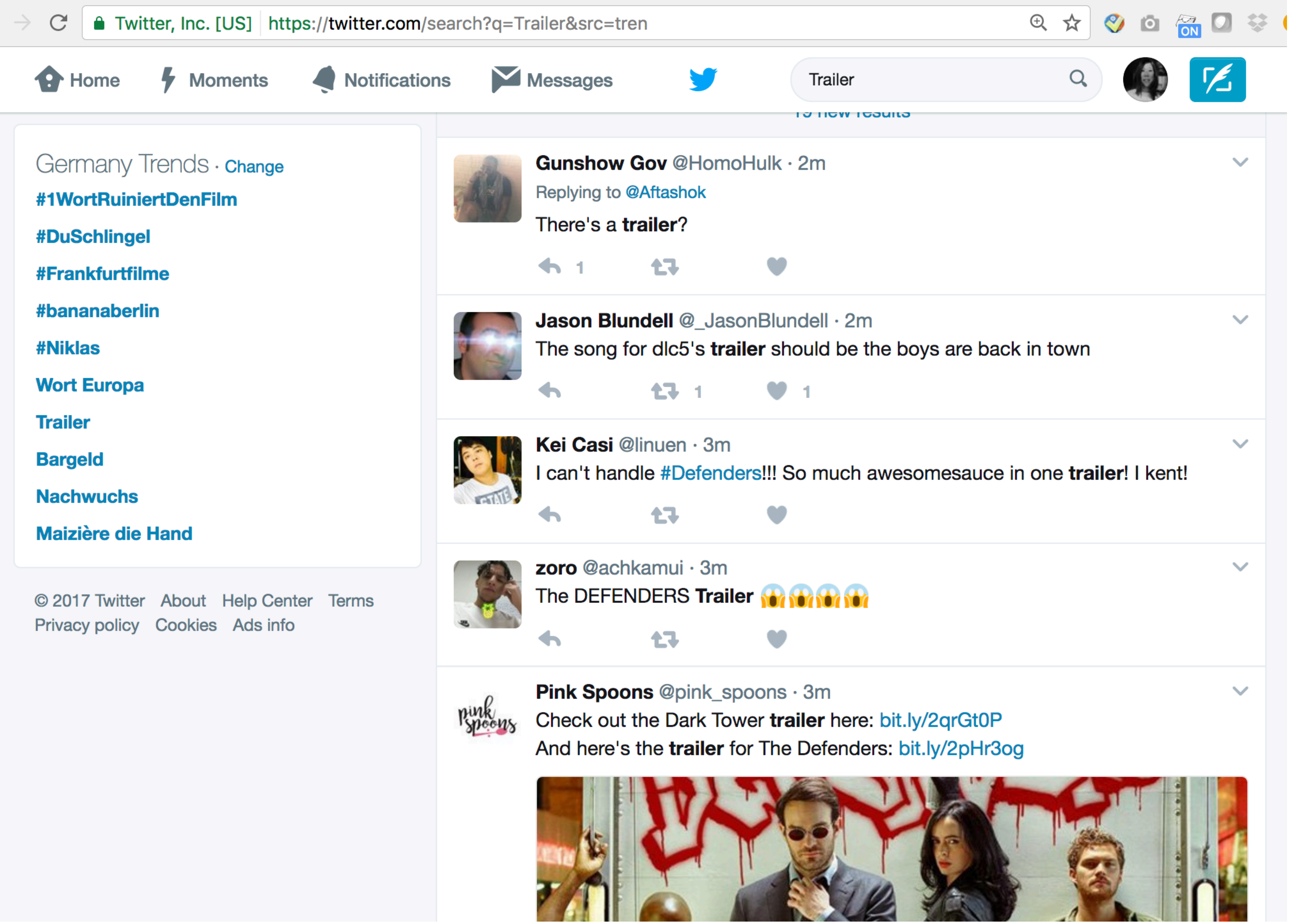
14,035 annotated pairs

needed human-in-the-loop to avoid "bad" topics

[1] Dolan et al., 2004
[2] Xu et al., 2014

Only exist two sentential paraphrase corpora

(which contain meaningful non-paraphrases)



[PIT-2015[2]]

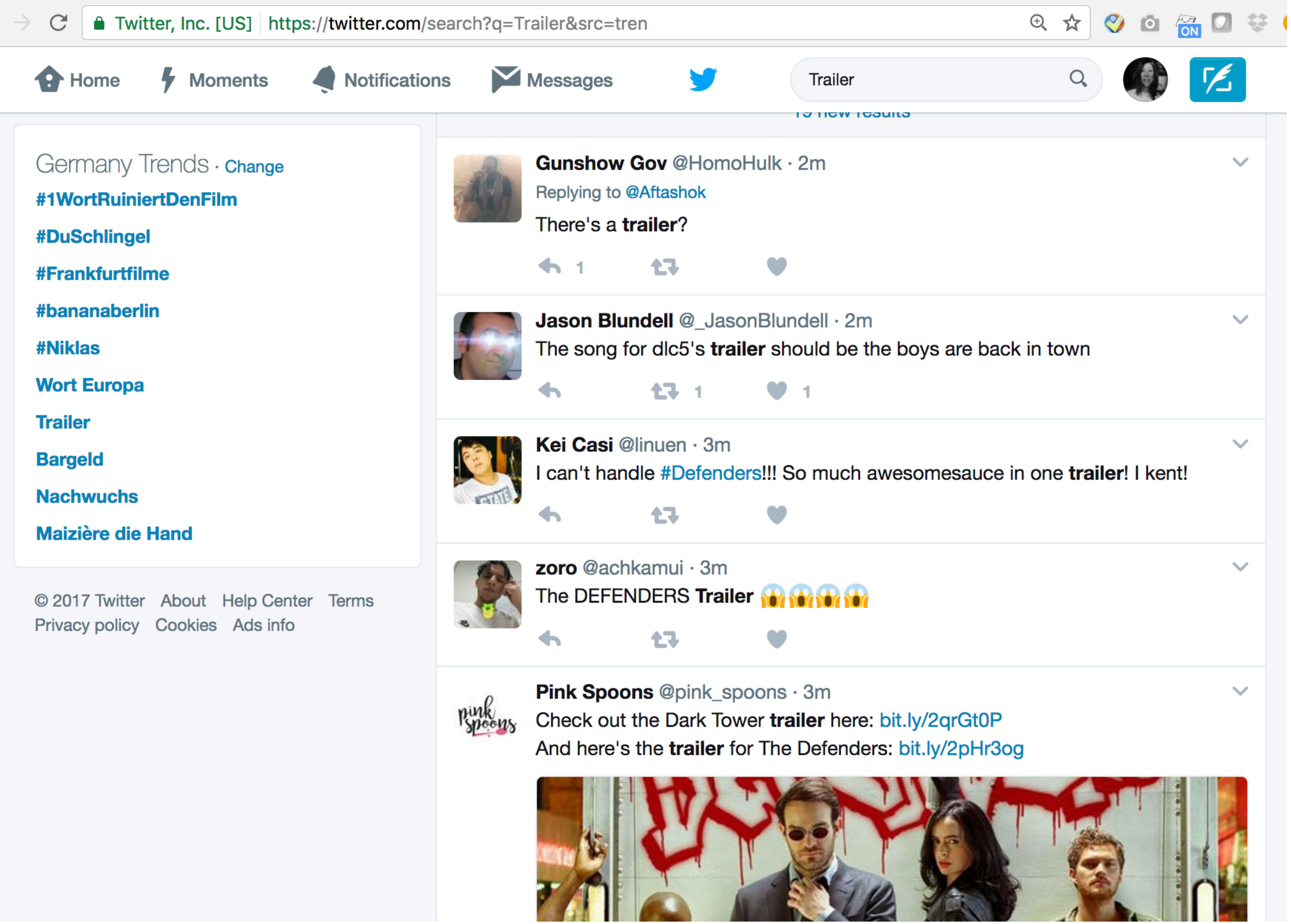
Twitter trending topics

14,035 annotated pairs

needed human-in-the-loop to avoid "bad" topics

Only exist two sentential paraphrase corpora

(which contain meaningful non-paraphrases)



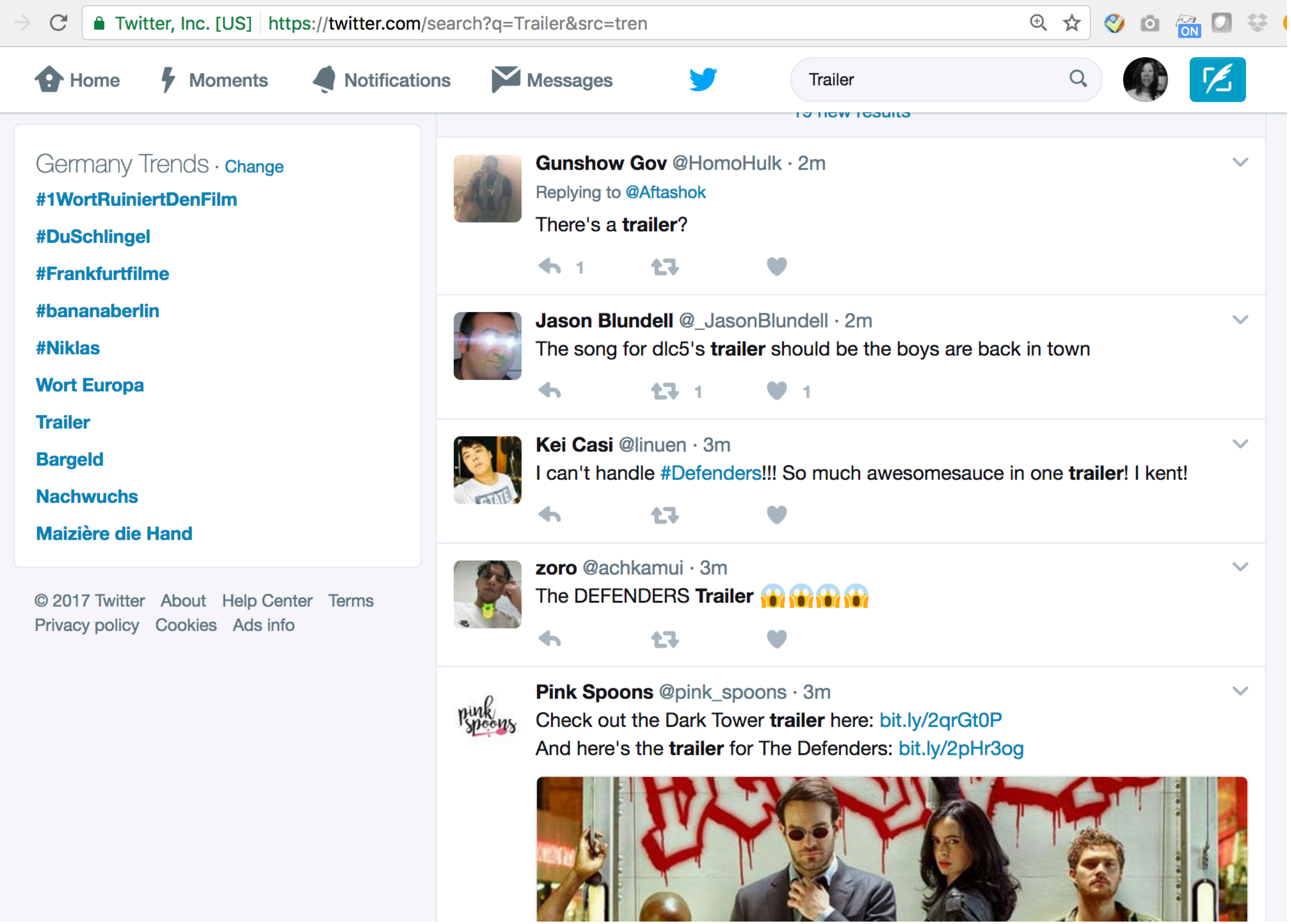
[PIT-2015[2]]

Twitter trending topics

14,035 annotated pairs

needed human-in-the-loop to avoid "bad" topics

Only exist two sentential paraphrase corpora
(which contain meaningful non-paraphrases)



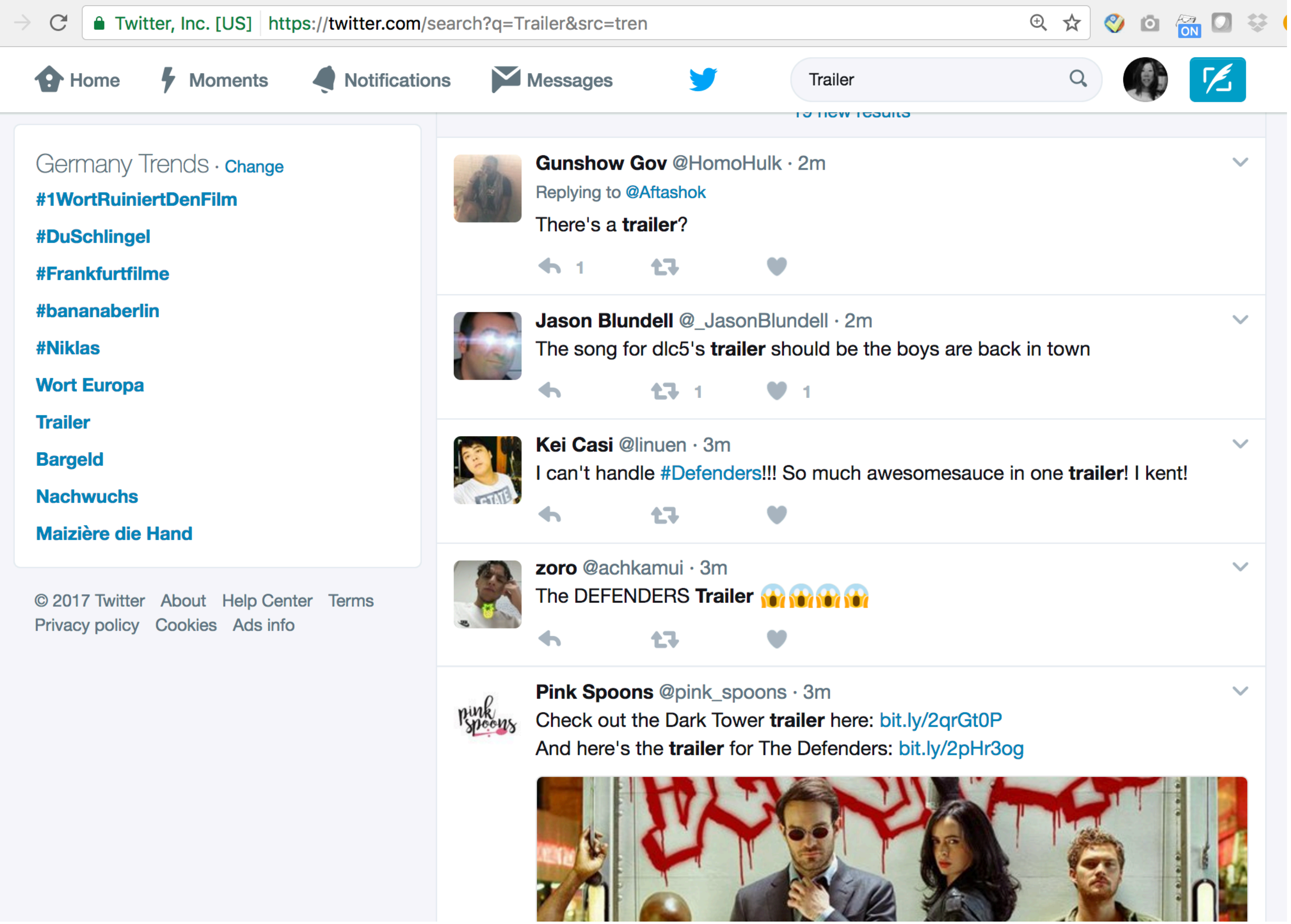
[PIT-2015[2]]

Twitter trending topics

14,035 annotated pairs

needed human-in-the-loop to avoid "bad" topics

Only exist two sentential paraphrase corpora
(which contain meaningful non-paraphrases)



[PIT-2015[2]]

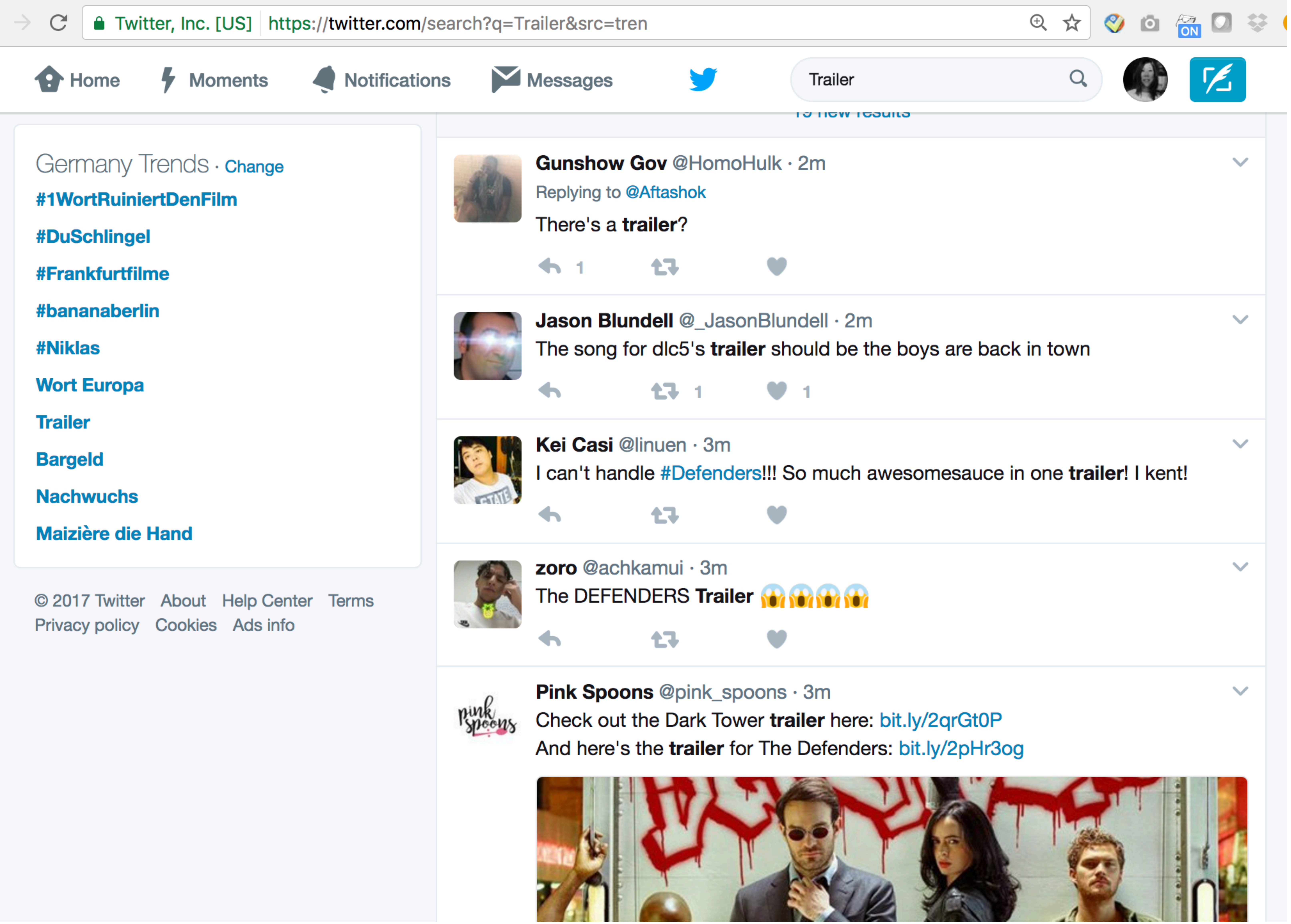
Twitter trending topics

14,035 annotated pairs

needed human-in-the-loop to avoid "bad" topics

Only exist two sentential paraphrase corpora

(which contain meaningful non-paraphrases)



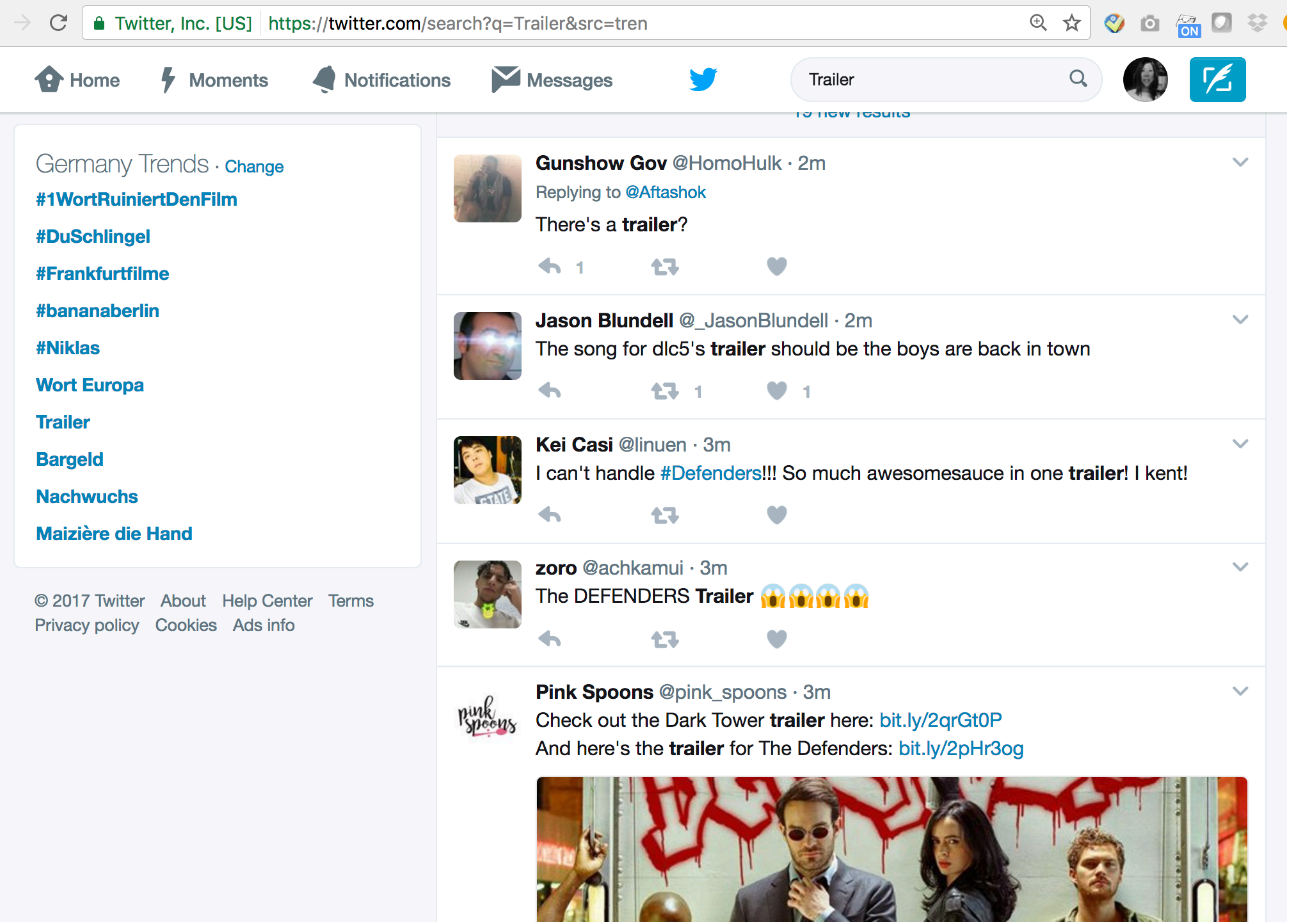
[PIT-2015[2]]

Twitter trending topics

14,035 annotated pairs

needed human-in-the-loop to avoid "bad" topics

Only exist two sentential paraphrase corpora
(which contain meaningful non-paraphrases)



[PIT-2015[2]]

Twitter trending topics

14,035 annotated pairs

needed human-in-the-loop to avoid “bad” topics

Only exist two sentential paraphrase corpora
(which contain meaningful non-paraphrases)

Key for success:

- narrow the search space
- ensure diversity among sentences

Also Pitfalls:

[MSRP[1]]

clustered news articles

5,801 annotated pairs

needed a SVM classifier to select sentences before data annotation

[PIT-2015[2]]

Twitter trending topics

14,035 annotated pairs

needed human-in-the-loop to avoid "bad" topics

[1] Dolan et al., 2004
[2] Xu et al., 2014

Only exist two sentential paraphrase corpora
(which contain meaningful non-paraphrases)

Key for success:

- narrow the search space
- ensure diversity among sentences

Also Pitfalls: cause over-identification when applied to unlabeled data

[MSRP[1]]

clustered news articles

5,801 annotated pairs

needed a SVM classifier to select sentences before data annotation

[PIT-2015[2]]

Twitter trending topics

14,035 annotated pairs

needed human-in-the-loop to avoid “bad” topics

[1] Dolan et al., 2004
[2] Xu et al., 2014

We created the 3rd paraphrase corpora

(largest annotated corpus to date)

Key for success:

- narrow the search space
- ensure diversity among sentences
- **the simpler the better!**

[MSRP[1]]

clustered
news articles

5,801 annotated pairs

[Twitter URL Corpus]

URL-linked
Tweets

51,524 annotated pairs

[PIT-2015[2]]

Twitter
trending topics

14,035 annotated pairs

[1] Dolan et al., 2004
[2] Xu et al., 2014

no clustering or topic detection needed
no data selection steps needed

We created the 3rd paraphrase corpora

(largest annotated corpus to date)

Key for success:

- narrow the search space
- ensure diversity among sentences
- **the simpler the better!**

[MSRP[1]]

clustered news articles

5,801 annotated pairs

[Twitter URL Corpus]

URL-linked Tweets

51,524 annotated pairs

largest up-to-date

[PIT-2015[2]]

Twitter trending topics

14,035 annotated pairs

[1] Dolan et al., 2004
[2] Xu et al., 2014

no clustering or topic detection needed
no data selection steps needed

We created the 3rd paraphrase corpora

(which also dynamically updates!)

Key for success:

- narrow the search space
- ensure diversity among sentences
- **the simpler the better! more effective automatic paraphrase identification**

[MSRP[1]]

clustered
news articles

5,801 annotated pairs

[Twitter URL Corpus]

URL-linked
Tweets

51,524 annotated pairs

**30,000 new sentential paraphrases
every month**

[PIT-2015[2]]

Twitter
trending topics

14,035 annotated pairs

[1] Dolan et al., 2004
[2] Xu et al., 2014

Once we have a lot of up-to-date sentential paraphrases

(we can, for example, learn name variations fully automatically)

Once we have a lot of up-to-date sentential paraphrases

(we can, for example, learn name variations fully automatically)

Donald Trump, DJT, Drumpf, Mr Trump, Idiot Trump, Chump, Evil Donald, #OrangeHitler, Donald @realTrump, D*nald Tr*mp, Comrade #Trump, Crooked #Trump, CryBaby Trump, Daffy Trump, Donald KKKrump, Dumb Trump, GOPTrump, Incompetent Trump, He-Who-Must-Not-Be-Named, Pres-elect Trump, President-Elect Trump, President-elect Donald J . Trump, PEOTUS Trump, Emperor Trump

Once we have a lot of up-to-date sentential paraphrases

(we can, of course, learn other synonyms in large quantity via word alignment)

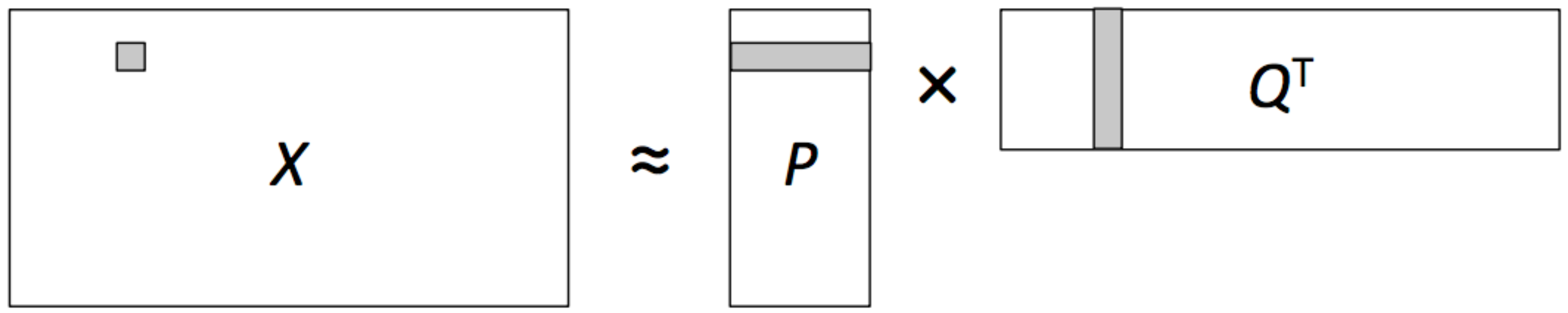
FBI Director backs CIA finding
FBI agrees with CIA
FBI backs CIA view
FBI finally backs CIA view
FBI now backs CIA view
FBI supports CIA assertion
FBI Clapper back CIA's view
The FBI backs the CIA's assessment
FBI Backs CIA …

How different from existing paraphrase corpora?

Model Performance → Dataset Difference

Automatic Paraphrase Identification

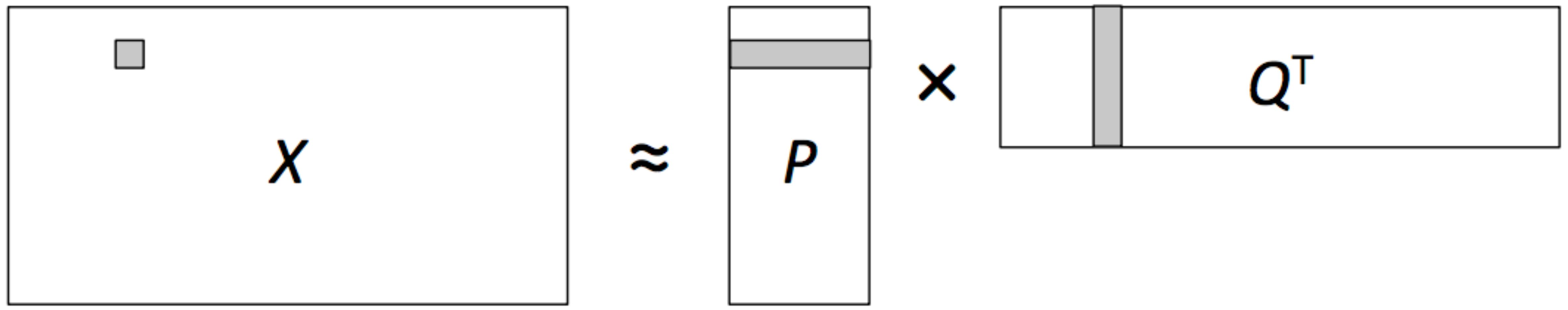
Automatic Paraphrase Identification



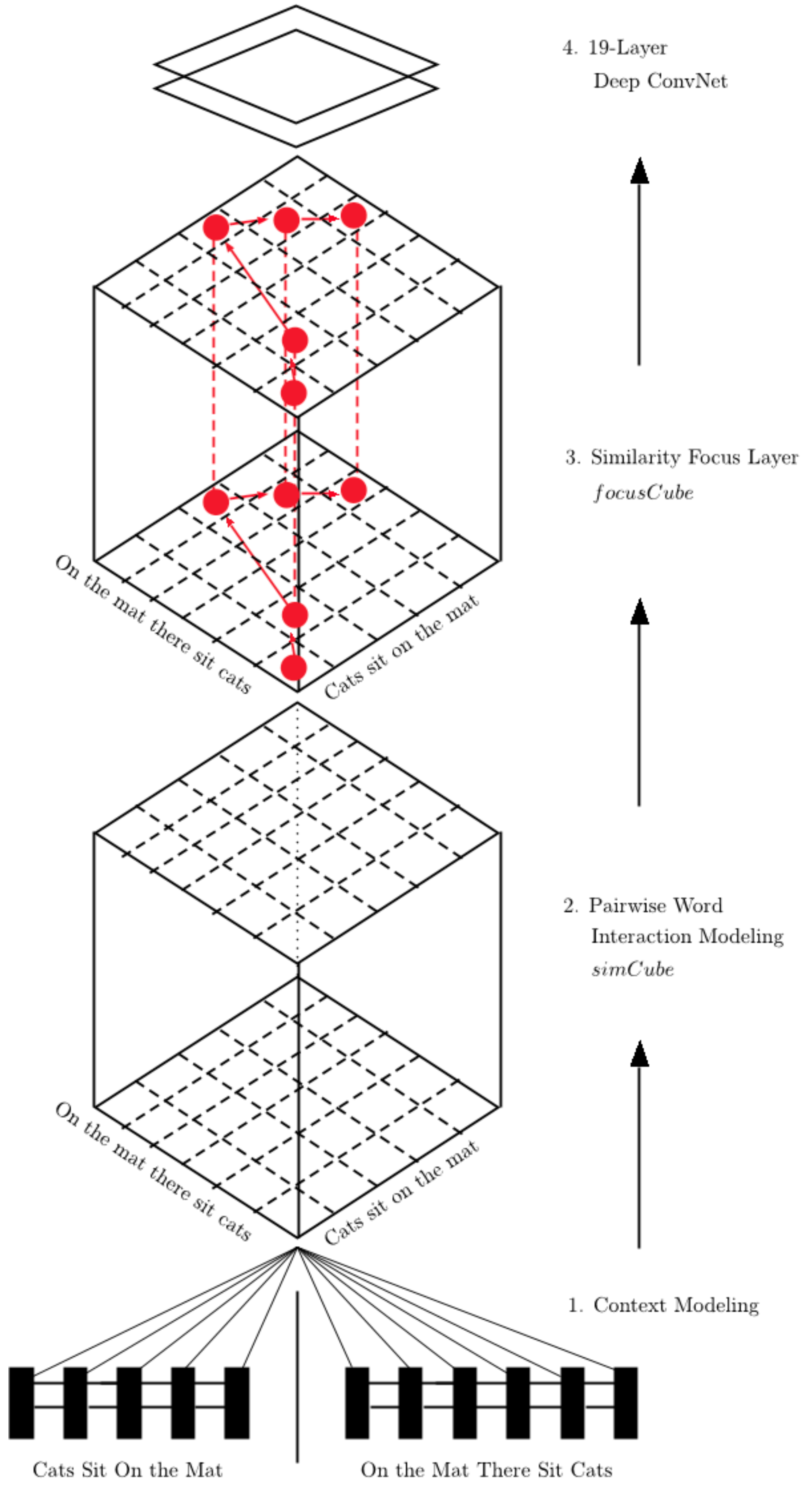
- **LEX-OrMF**[1] (Orthogonal Matrix Factorization[2])

[1] Xu et al., 2014
[2] Guo et al., 2014

Automatic Paraphrase Identification

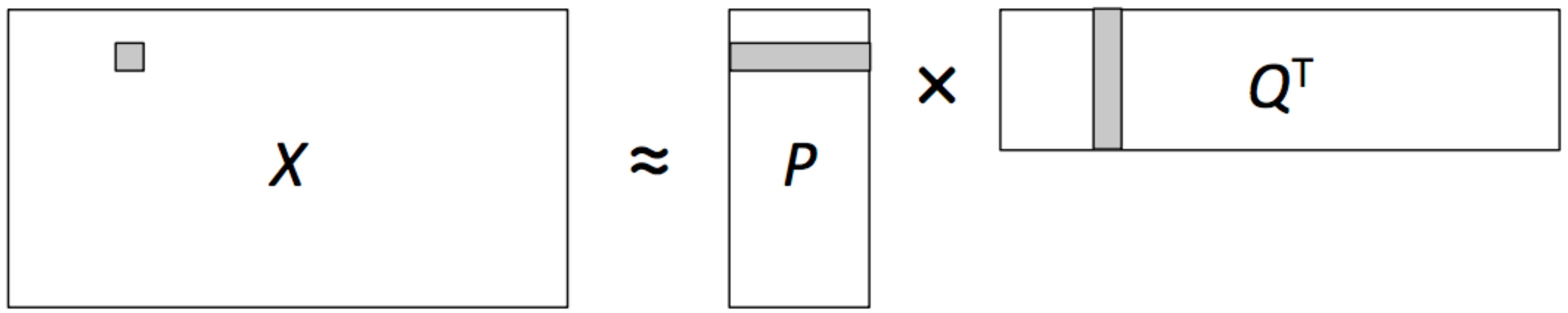


- **LEX-OrMF**[1] (Orthogonal Matrix Factorization[2])
- **DeepPairwiseWord**[3] (Deep Neural Networks)



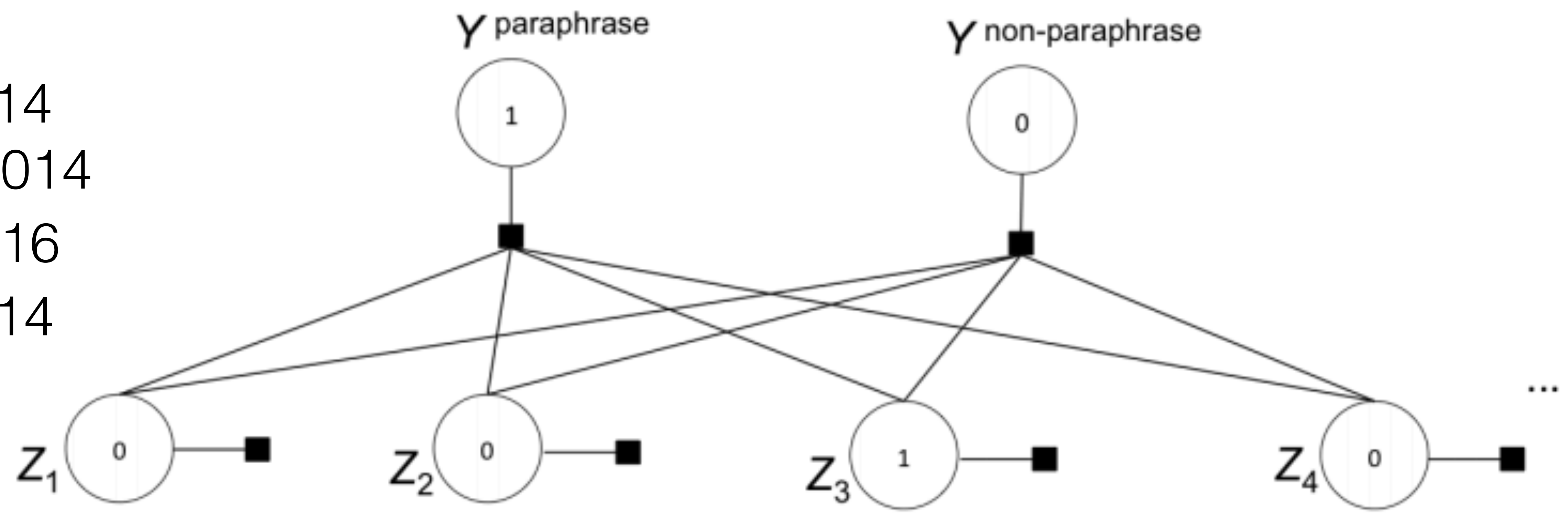
[1] Xu et al., 2014
[2] Guo et al., 2014
[3] He et al., 2016

Automatic Paraphrase Identification

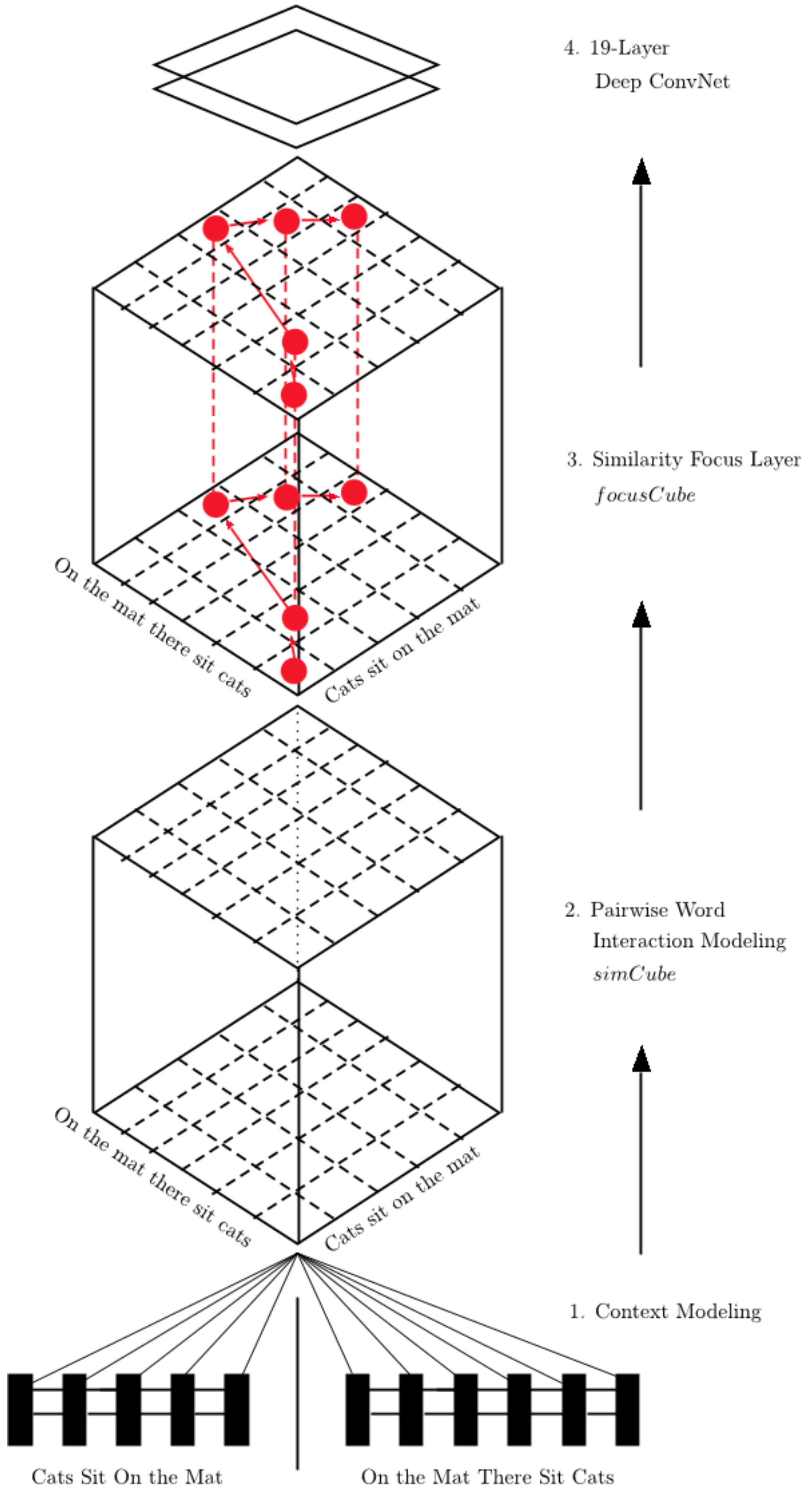


- **LEX-OrMF**[1] (Orthogonal Matrix Factorization[2])
- **DeepPairwiseWord**[3] (Deep Neural Networks)
- **MultiP**[4] (Multiple Instance Learning)

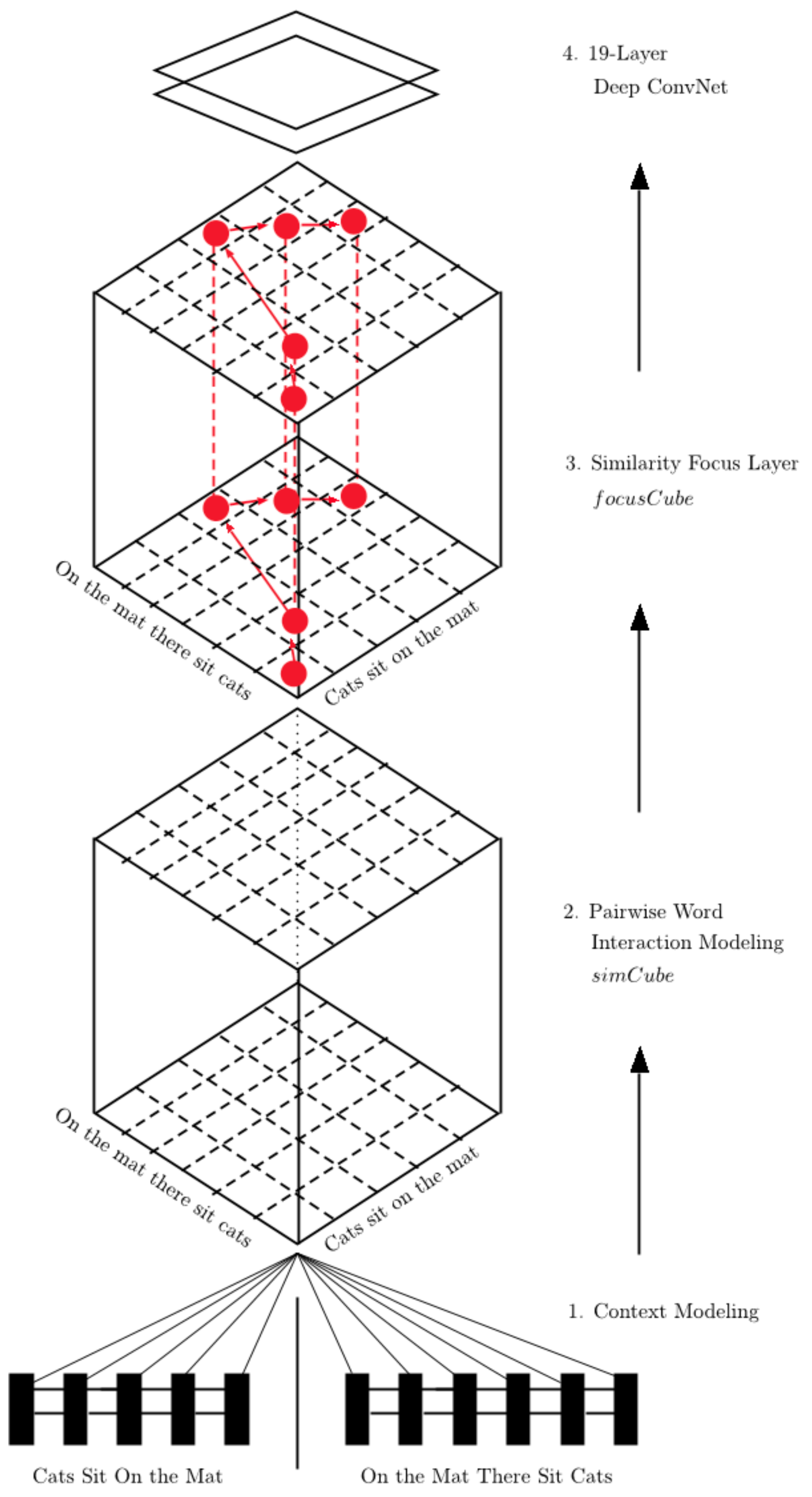
[1] Xu et al., 2014
[2] Guo et al., 2014
[3] He et al., 2016
[4] Xu et al., 2014



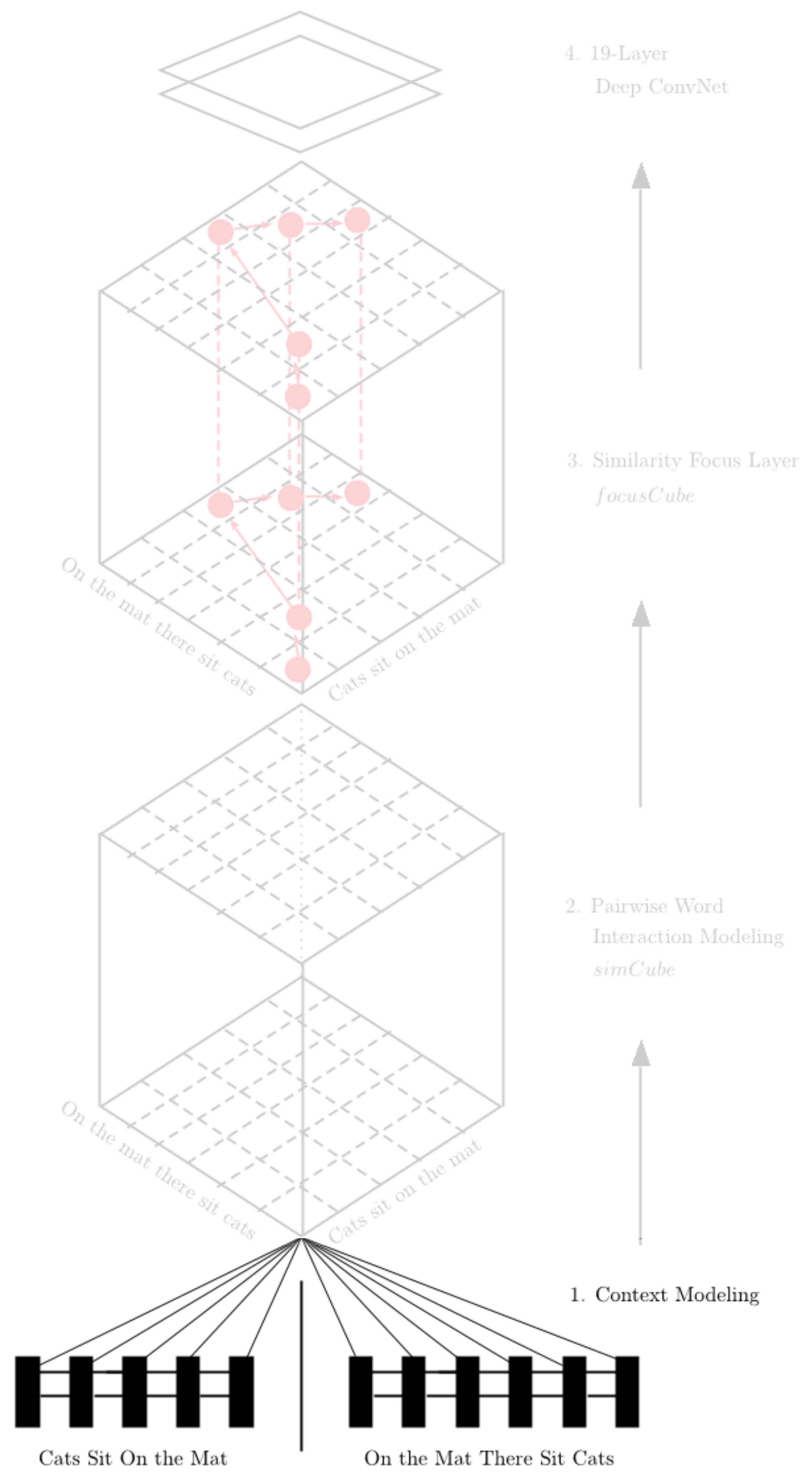
$$P(\mathbf{z}_i, y_i | \mathbf{w}_i; \theta) = \prod_{j=1}^{m} \exp(\theta \cdot f(z_j, w_j)) \times \sigma(\mathbf{z}_i, y_i)$$



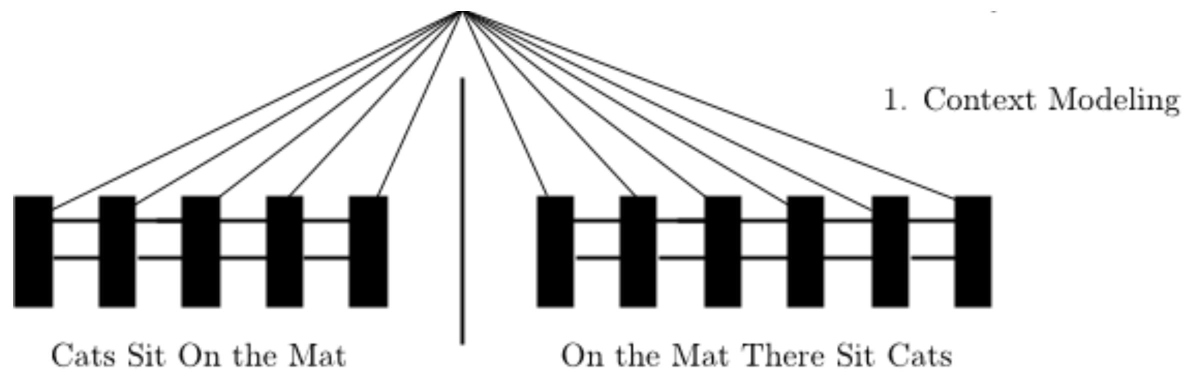
Deep Pairwise Word Model



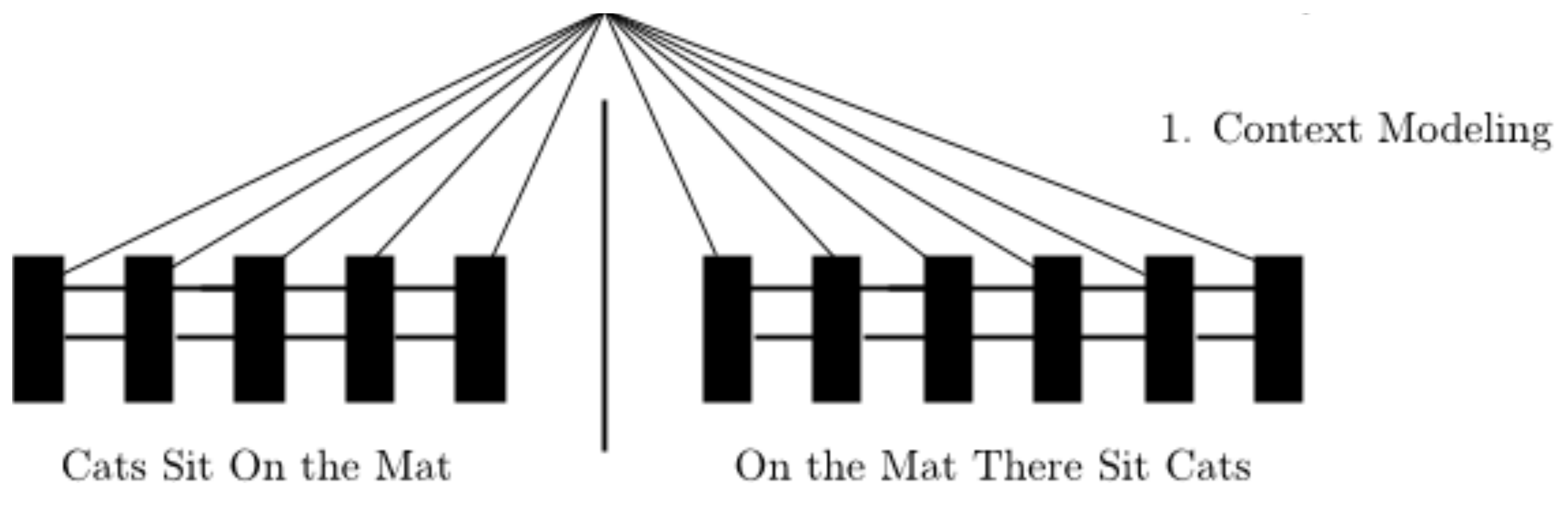
Deep Pairwise Word Model



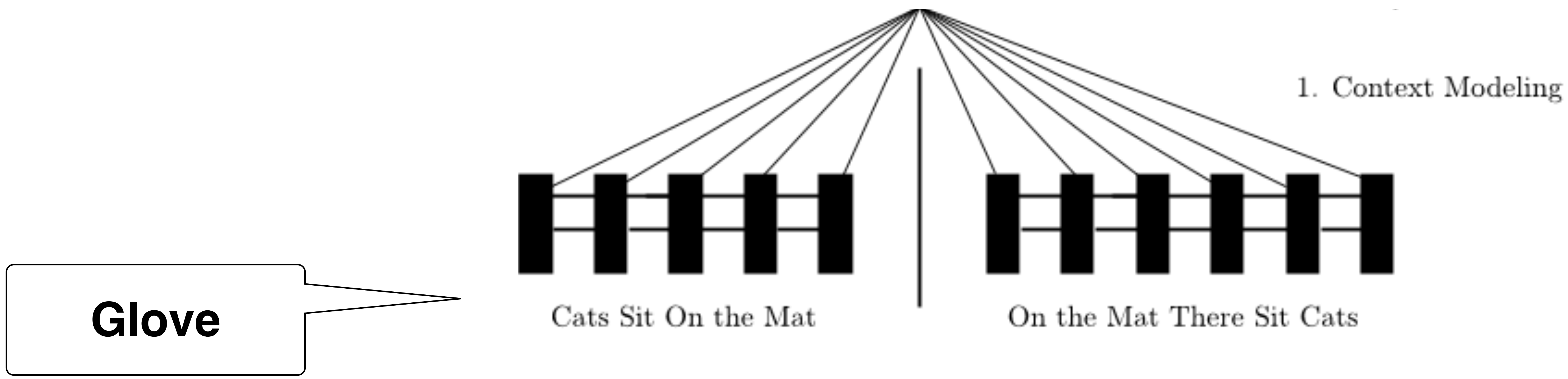
Deep Pairwise Word Model



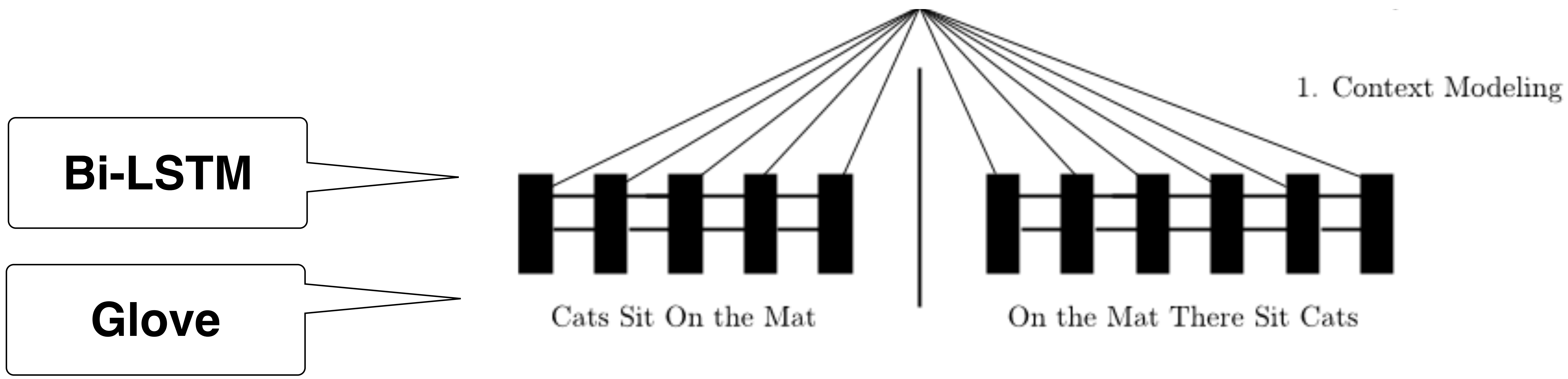
Deep Pairwise Word Model



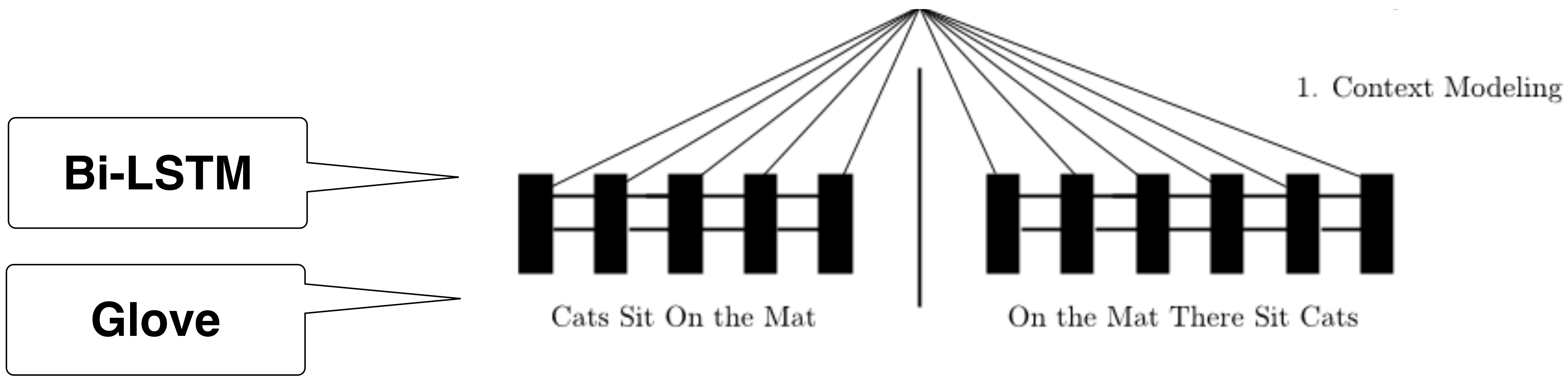
Deep Pairwise Word Model



Deep Pairwise Word Model

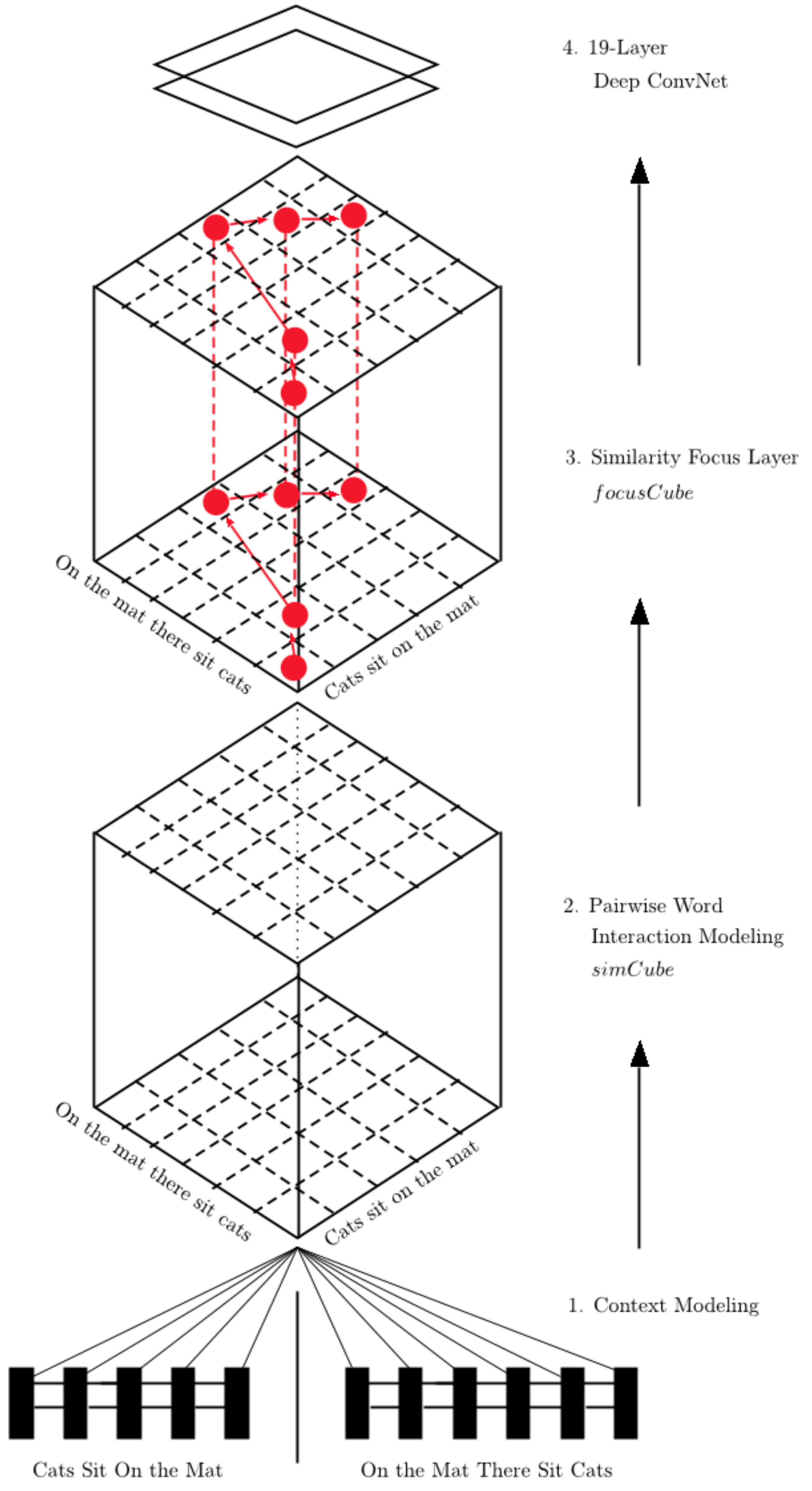


Deep Pairwise Word Model

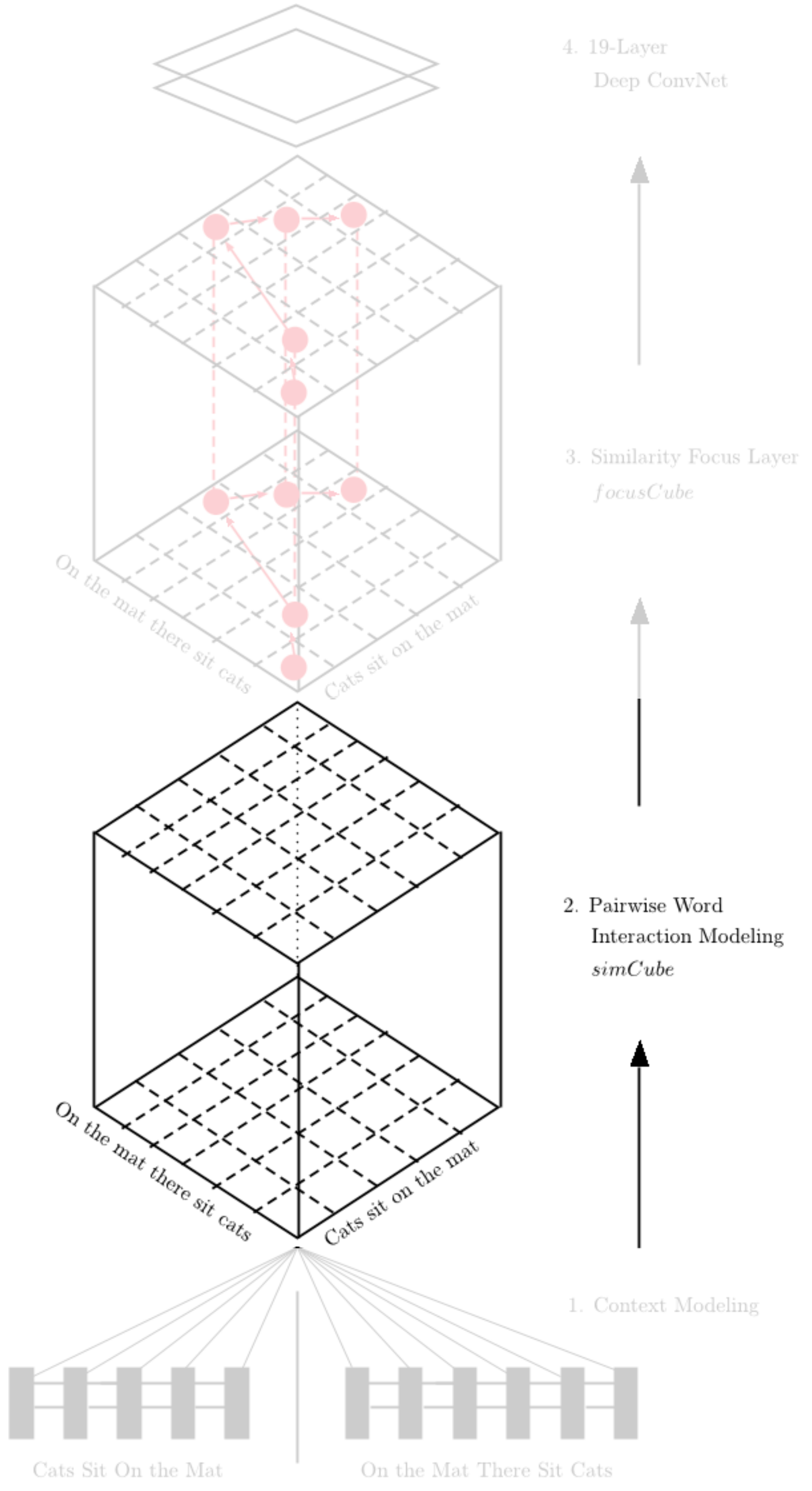


Decompose sentence input into word context to reduce modeling difficulty

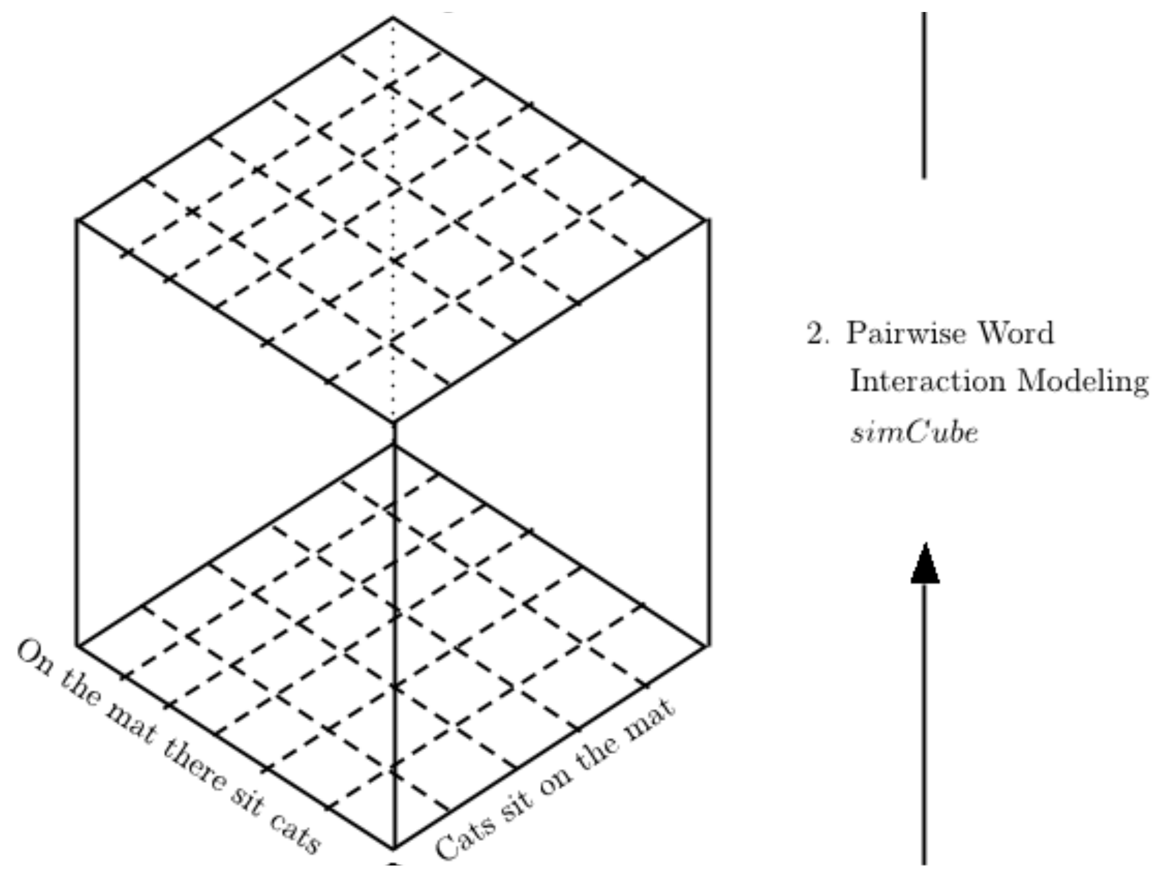
Deep Pairwise Word Model



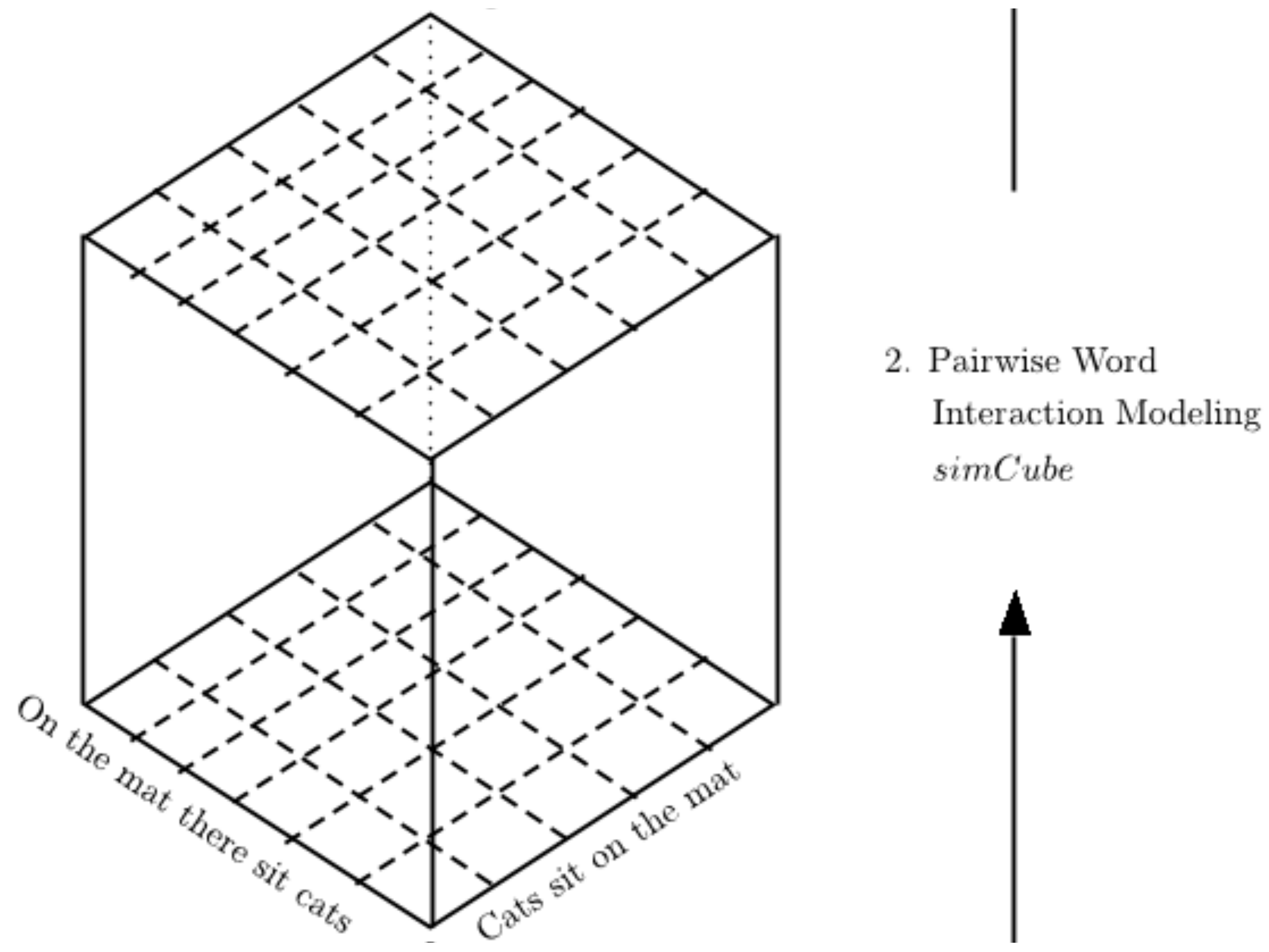
Deep Pairwise Word Model



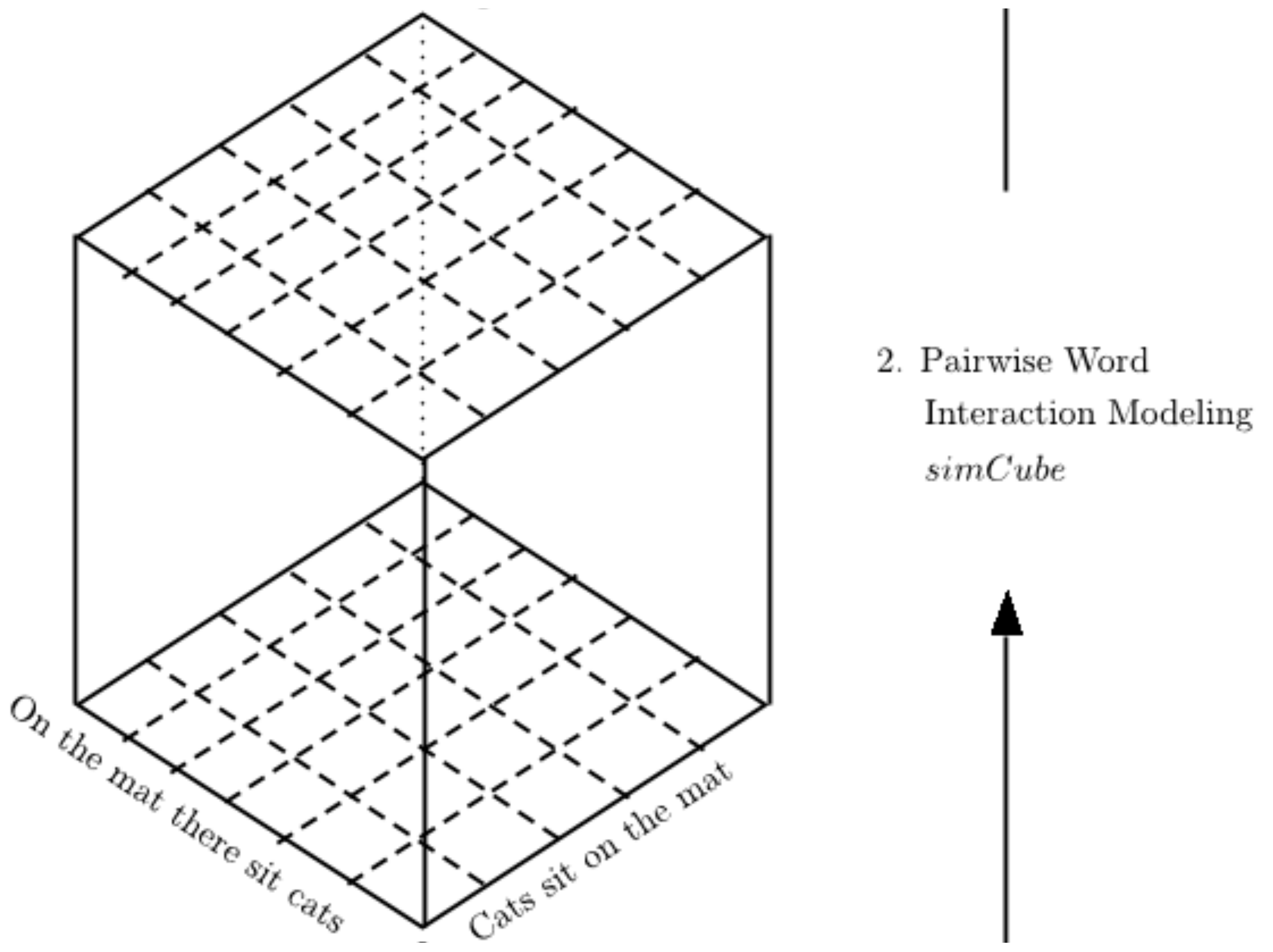
Deep Pairwise Word Model



Deep Pairwise Word Model

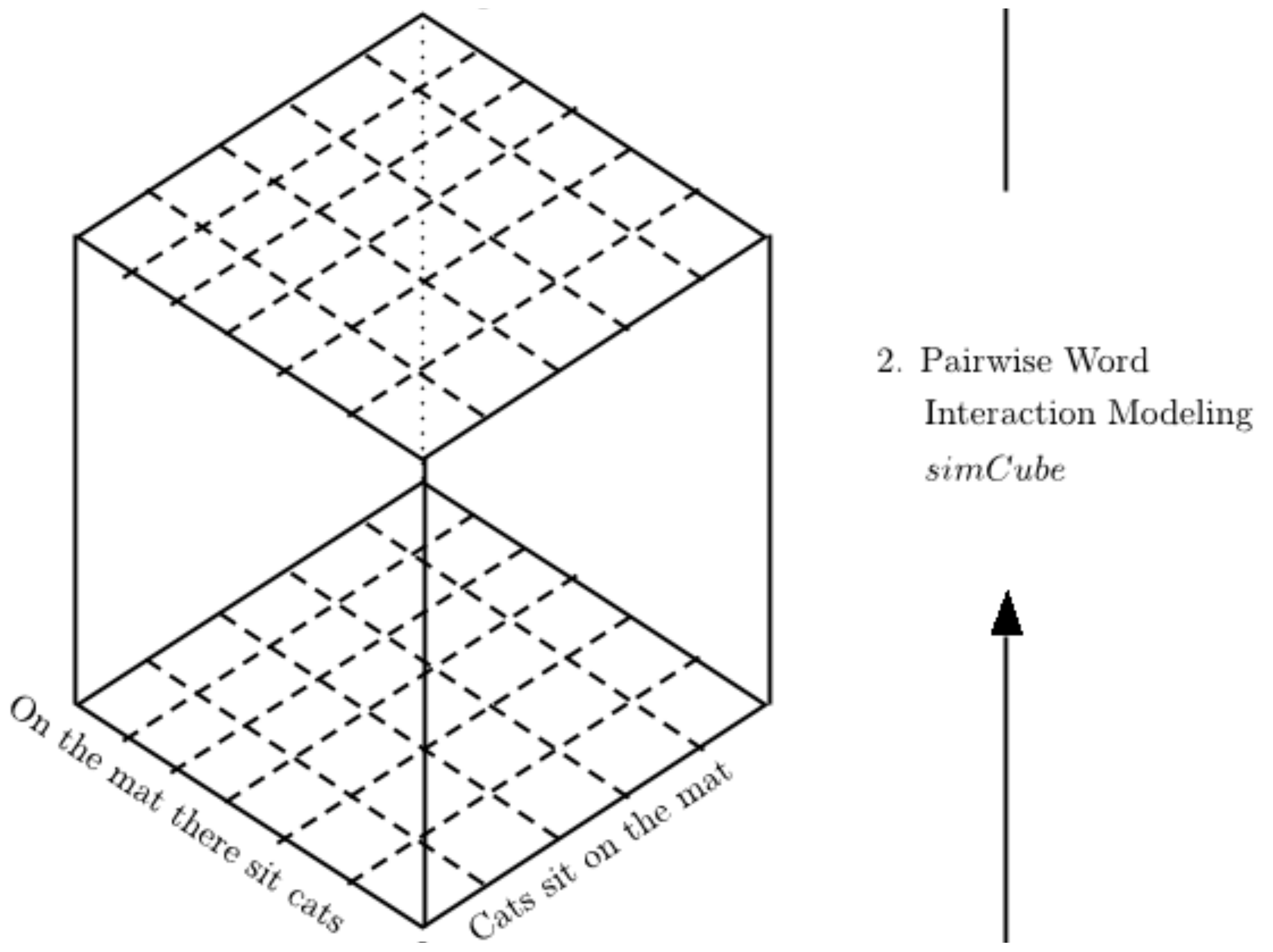


Deep Pairwise Word Model



$$coU(\overrightarrow{h_1}, \overrightarrow{h_2}) = \{\cos(\overrightarrow{h_1}, \overrightarrow{h_2}), L_2Euclid(\overrightarrow{h_1}, \overrightarrow{h_2}), DotProduct(\overrightarrow{h_1}, \overrightarrow{h_2})\}$$

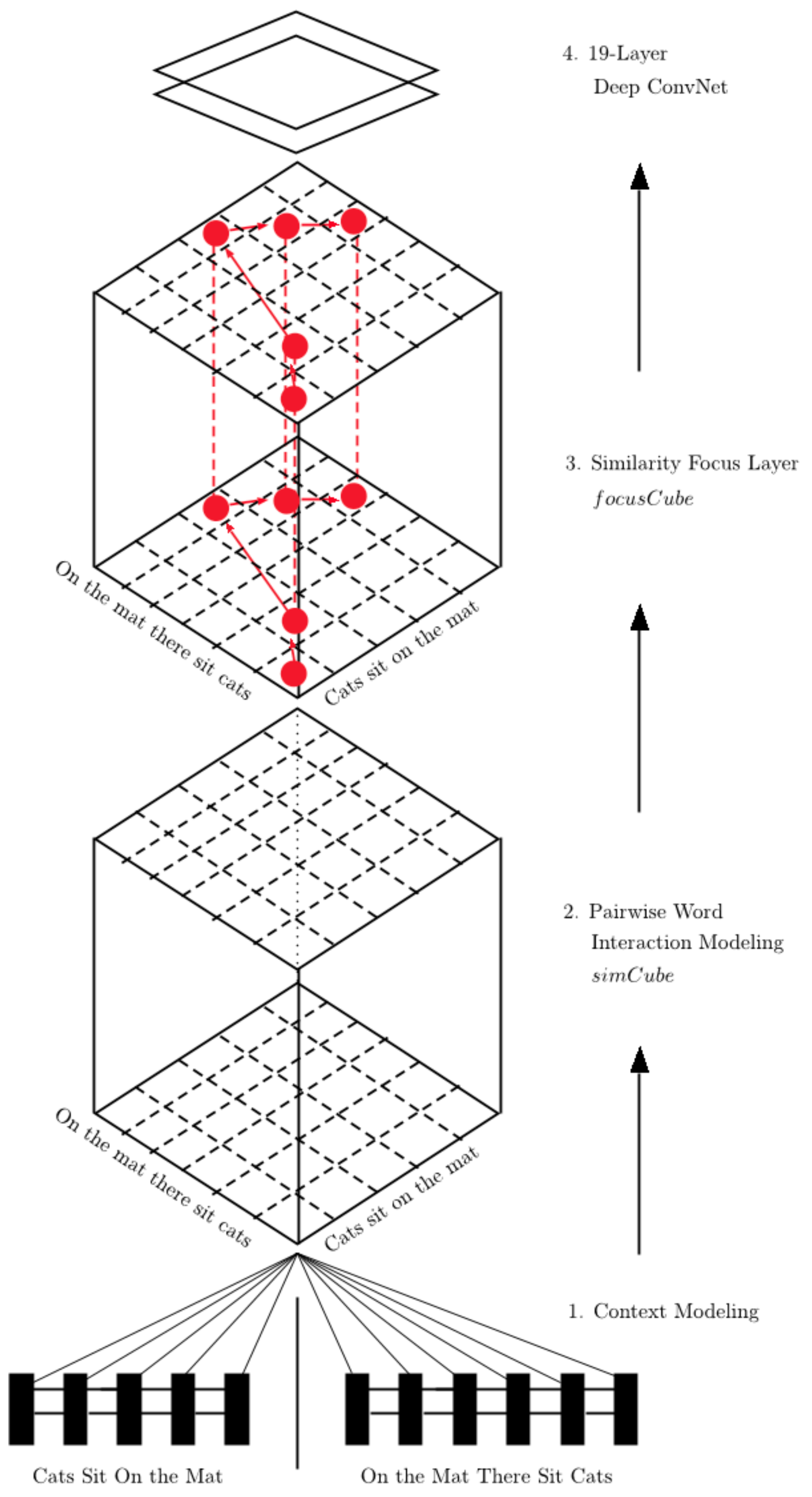
Deep Pairwise Word Model



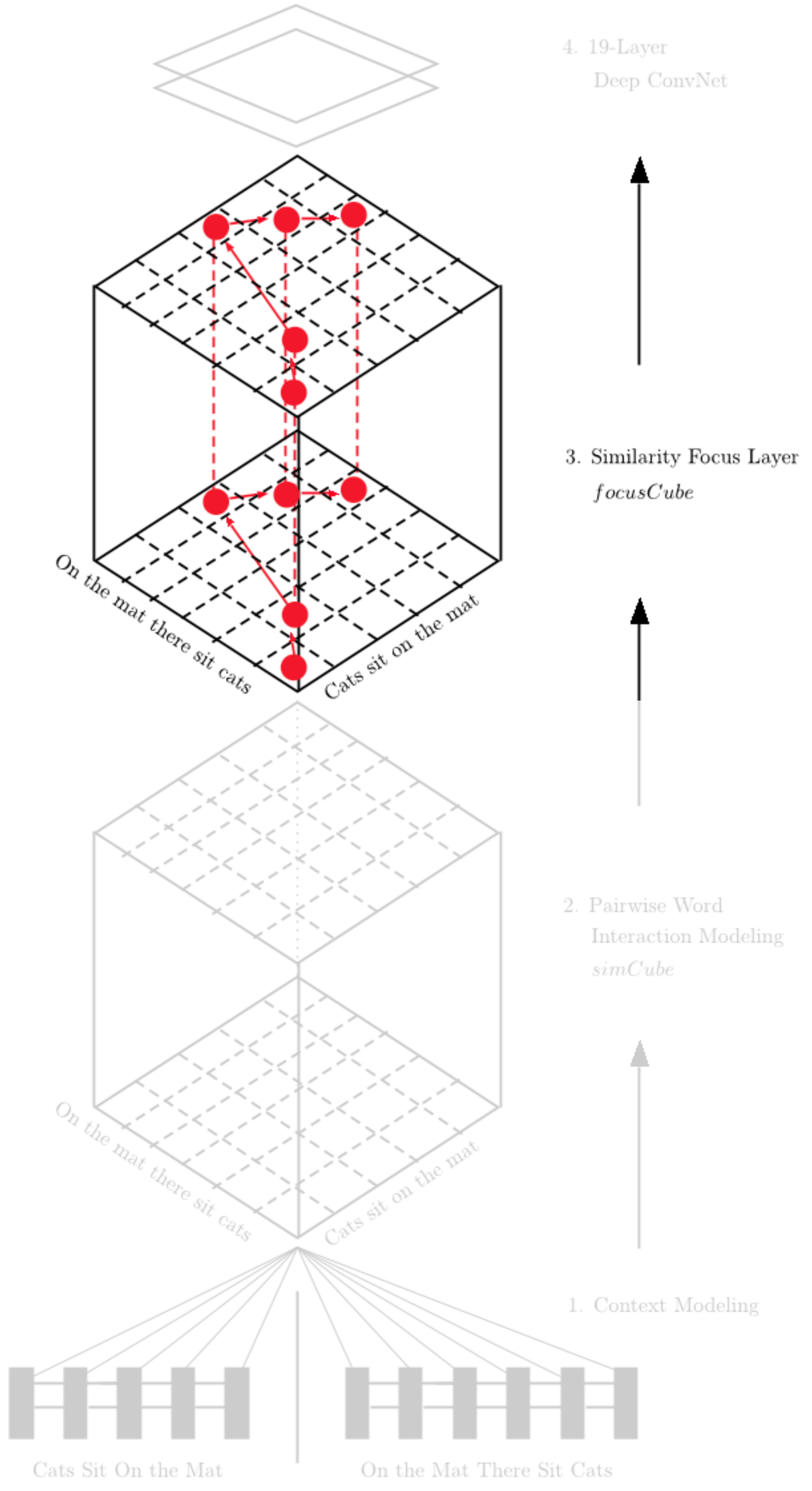
$$coU(\overrightarrow{h_1}, \overrightarrow{h_2}) = \{\cos(\overrightarrow{h_1}, \overrightarrow{h_2}), L_2 Euclid(\overrightarrow{h_1}, \overrightarrow{h_2}), DotProduct(\overrightarrow{h_1}, \overrightarrow{h_2})\}$$

Multiple vector similarity measurement used to capture word pair relationship

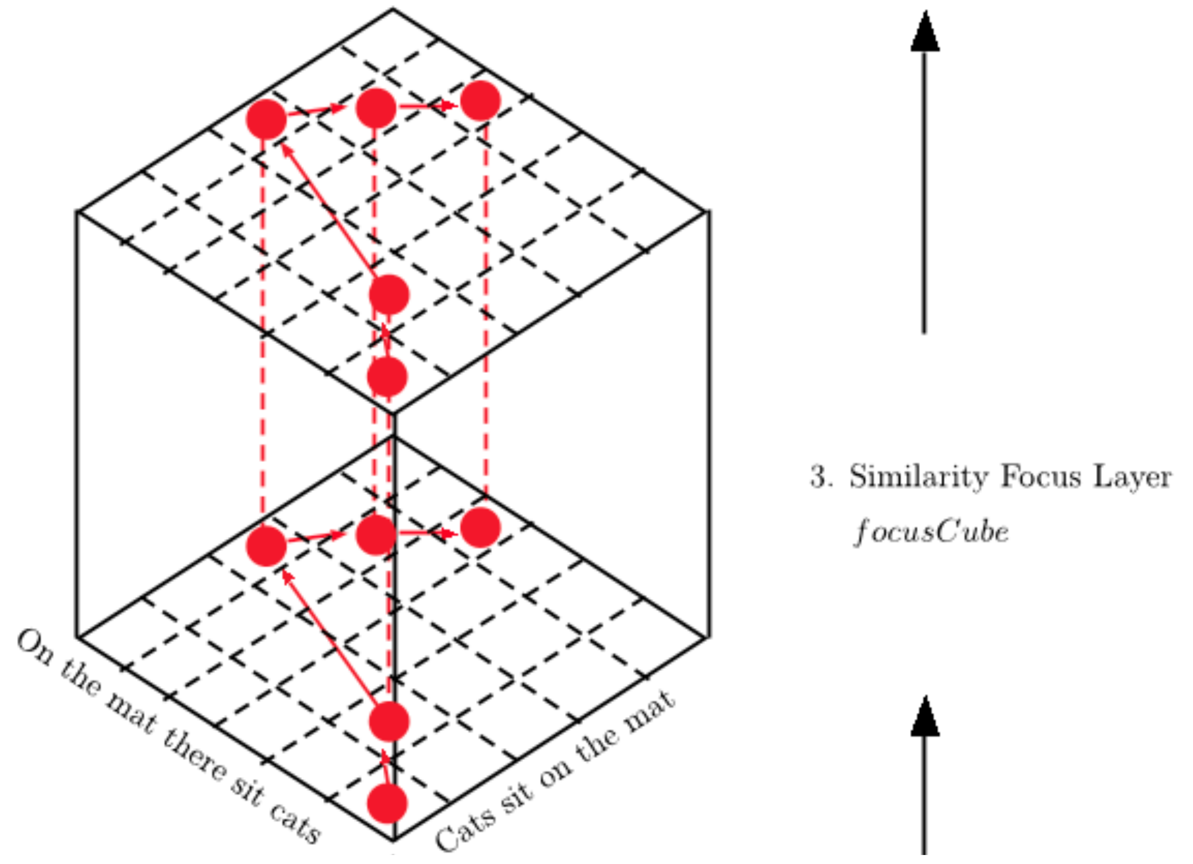
Deep Pairwise Word Model



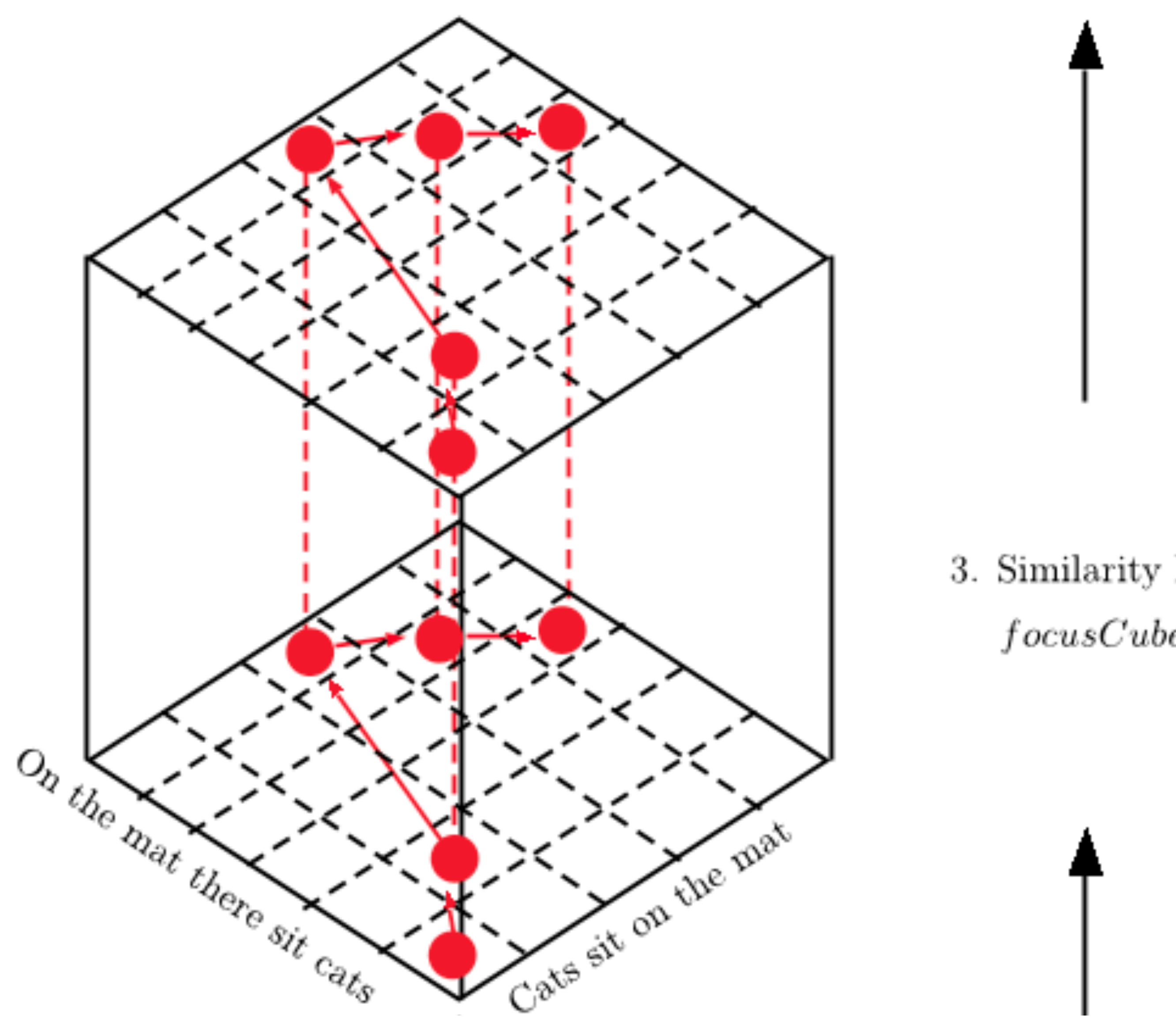
Deep Pairwise Word Model



Deep Pairwise Word Model



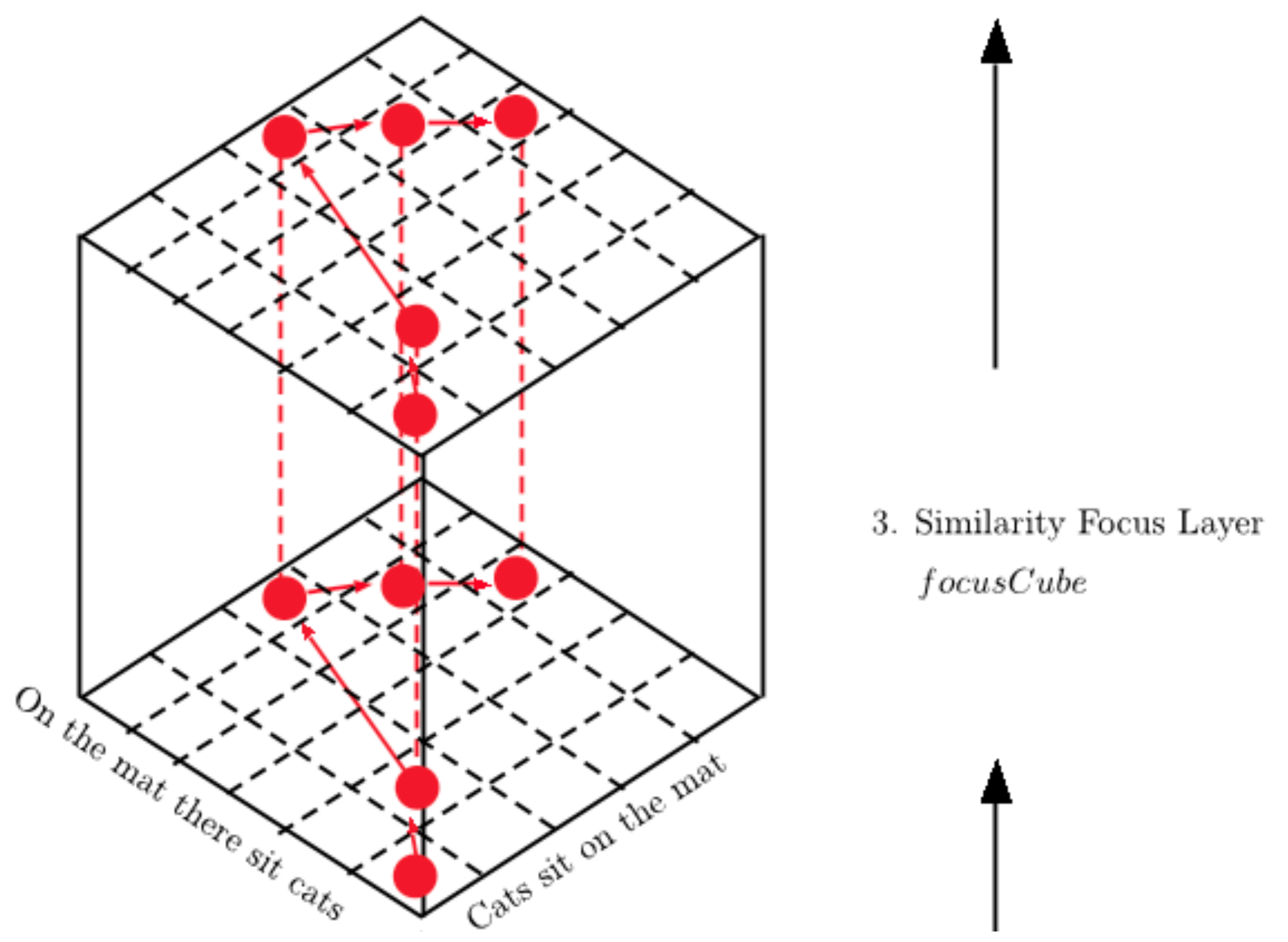
Deep Pairwise Word Model



3. Similarity Focus Layer

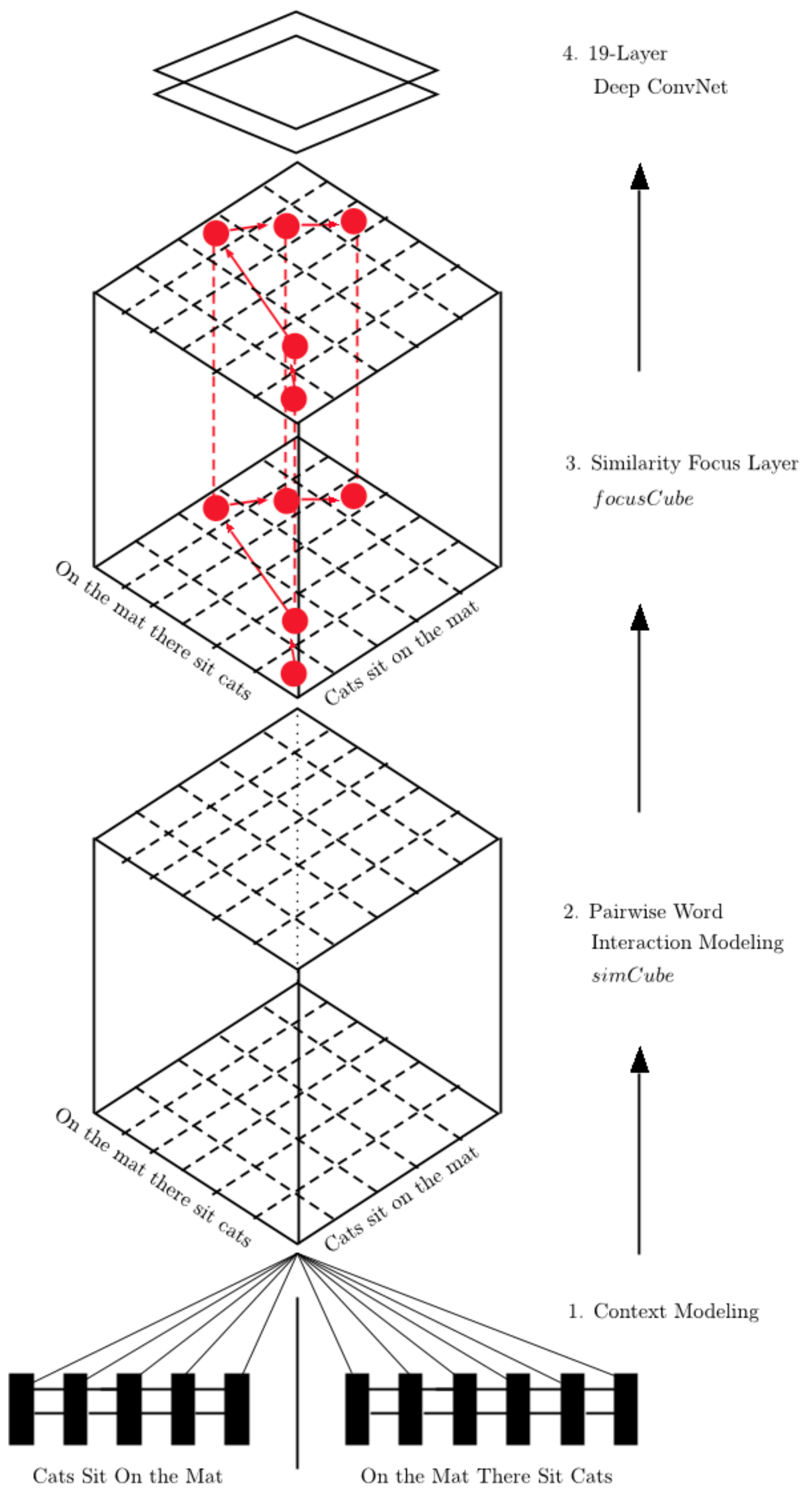
$focusCube$

Deep Pairwise Word Model

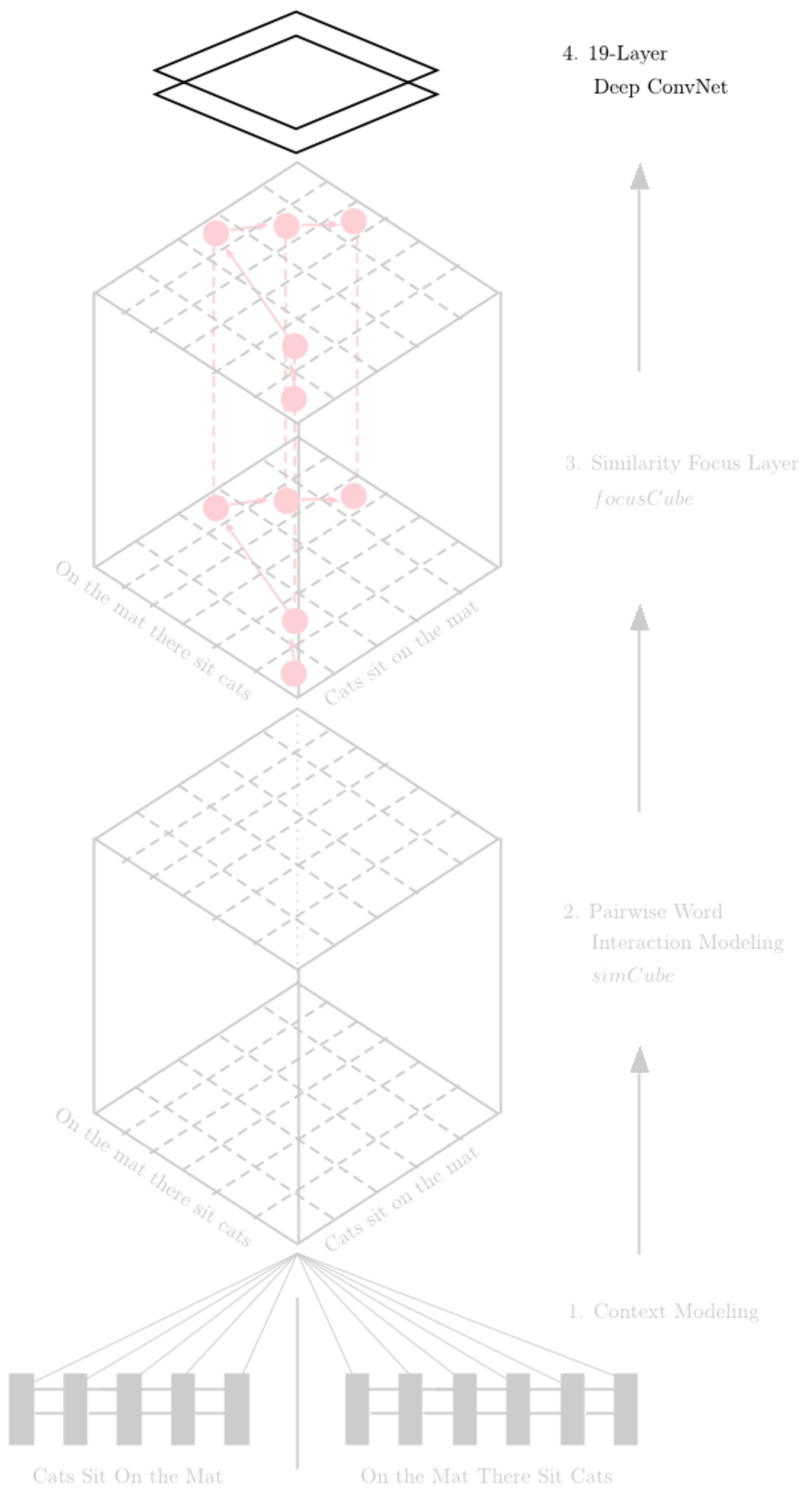


More attention added to top ranked word pairs.

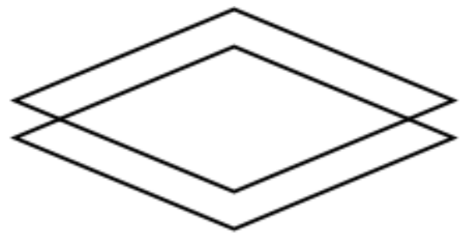
Deep Pairwise Word Model



Deep Pairwise Word Model

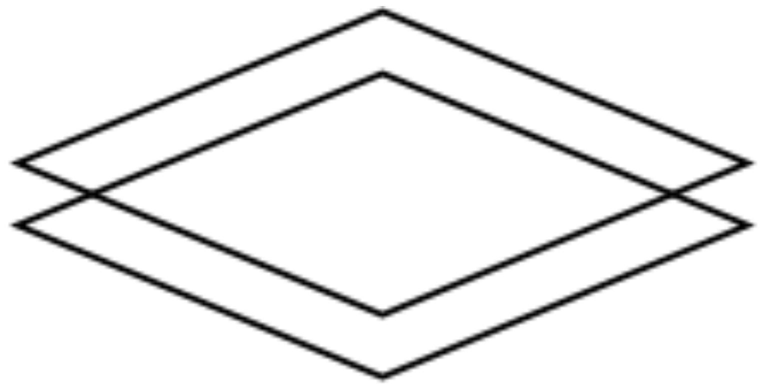


Deep Pairwise Word Model



4. 19-Layer
Deep ConvNet

Deep Pairwise Word Model

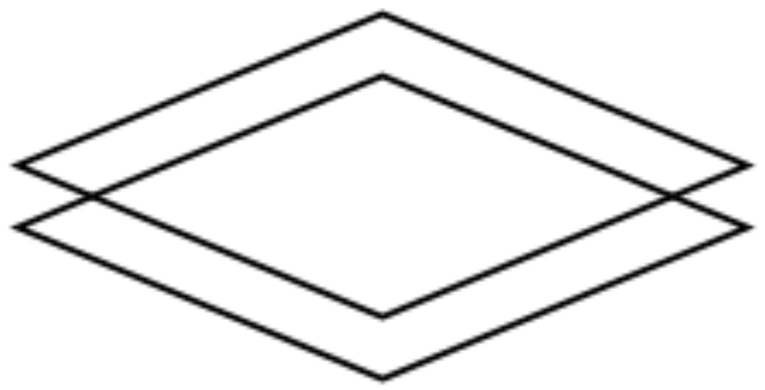


4. 19-Layer
   Deep ConvNet

Deep Pairwise Word Model

Deep ConvNet Configurations	
Input Size: 32 **by** 32	**Input Size:** 48 **by** 48
Spatial Conv 128: size 3×3, stride 1, pad 1	
ReLU	
Max Pooling: size 2×2, stride 2	
Spatial Conv 164: size 3×3, stride 1, pad 1	
ReLU	
Max Pooling: size 2×2, stride 2	
Spatial Conv 192: size 3×3, stride 1, pad 1	
ReLU	
Max Pooling: size 2×2, stride 2	
Spatial Conv 192: size 3×3, stride 1, pad 1	
ReLU	
Max Pooling: size 2×2, stride 2	
Spatial Conv 128: size 3×3, stride 1, pad 1	
ReLU	
Max Pooling: 2×2, s2	Max Pooling: 3×3, s1
Fully-Connected Layer	
ReLU	
Fully-Connected Layer	
LogSoftMax	

Table 1: Deep ConvNet architecture given two padding size configurations for final classification.

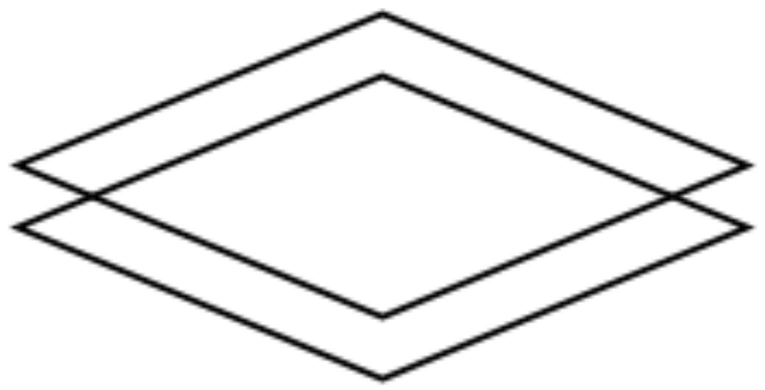


4. 19-Layer Deep ConvNet

Deep Pairwise Word Model

Deep ConvNet Configurations	
Input Size: 32 **by** 32	**Input Size:** 48 **by** 48
Spatial Conv 128: size 3×3, stride 1, pad 1	
ReLU	
Max Pooling: size 2×2, stride 2	
Spatial Conv 164: size 3×3, stride 1, pad 1	
ReLU	
Max Pooling: size 2×2, stride 2	
Spatial Conv 192: size 3×3, stride 1, pad 1	
ReLU	
Max Pooling: size 2×2, stride 2	
Spatial Conv 192: size 3×3, stride 1, pad 1	
ReLU	
Max Pooling: size 2×2, stride 2	
Spatial Conv 128: size 3×3, stride 1, pad 1	
ReLU	
Max Pooling: 2×2, s2	Max Pooling: 3×3, s1
Fully-Connected Layer	
ReLU	
Fully-Connected Layer	
LogSoftMax	

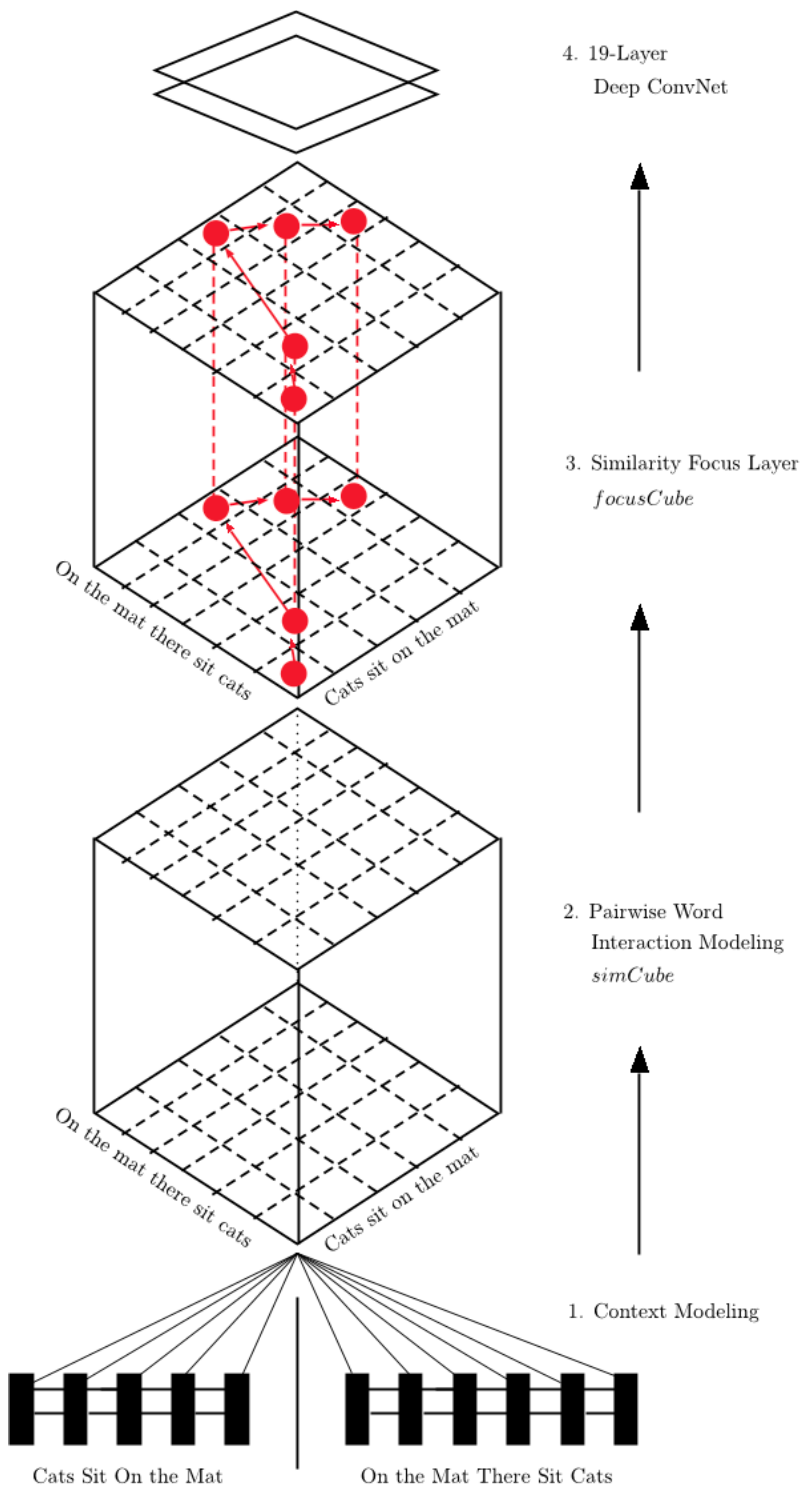
Table 1: Deep ConvNet architecture given two padding size configurations for final classification.



4. 19-Layer Deep ConvNet

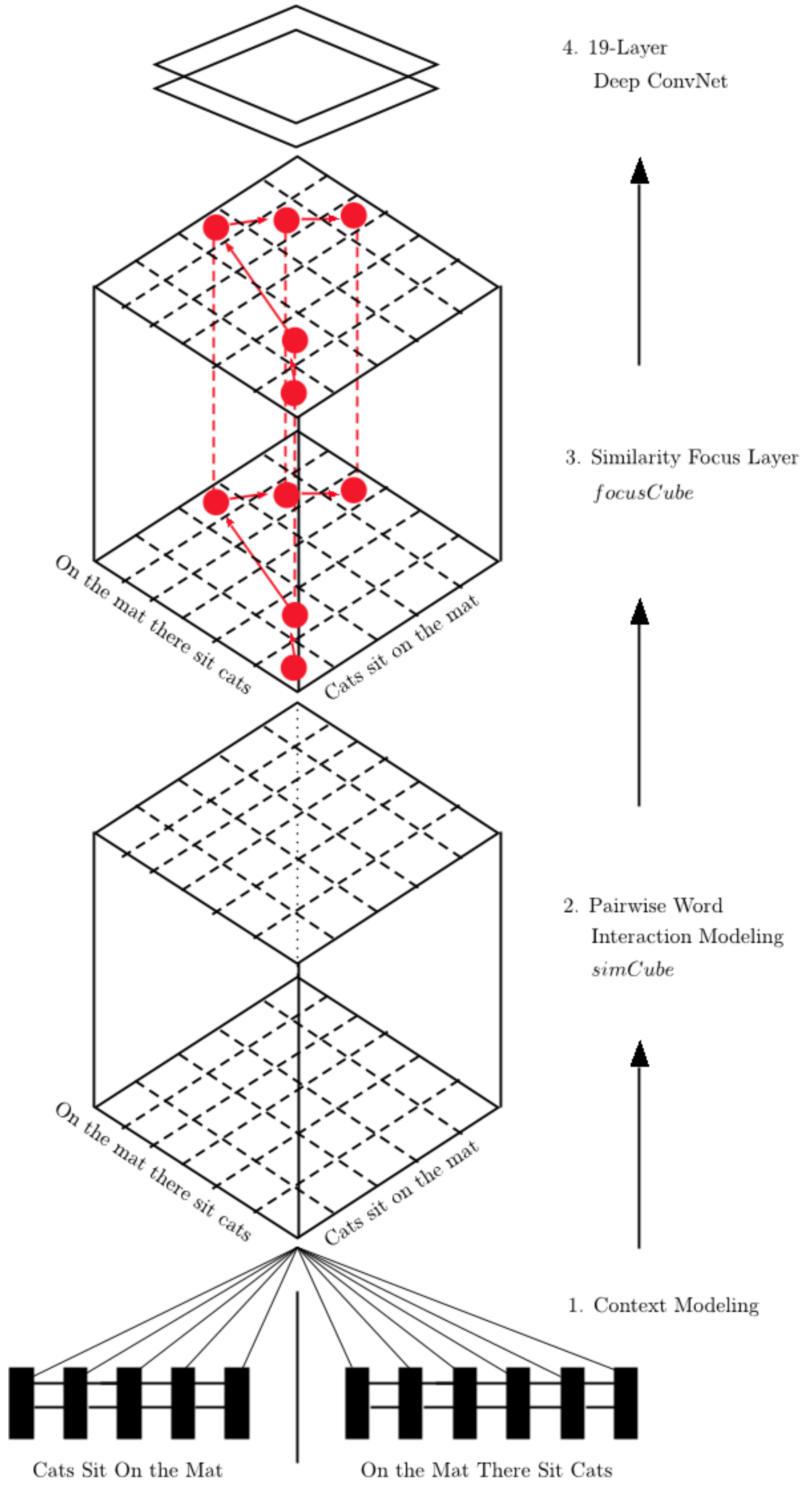
Sentence pair relationship can be identified by pattern recognition through ConvNet.

Deep Pairwise Word Model



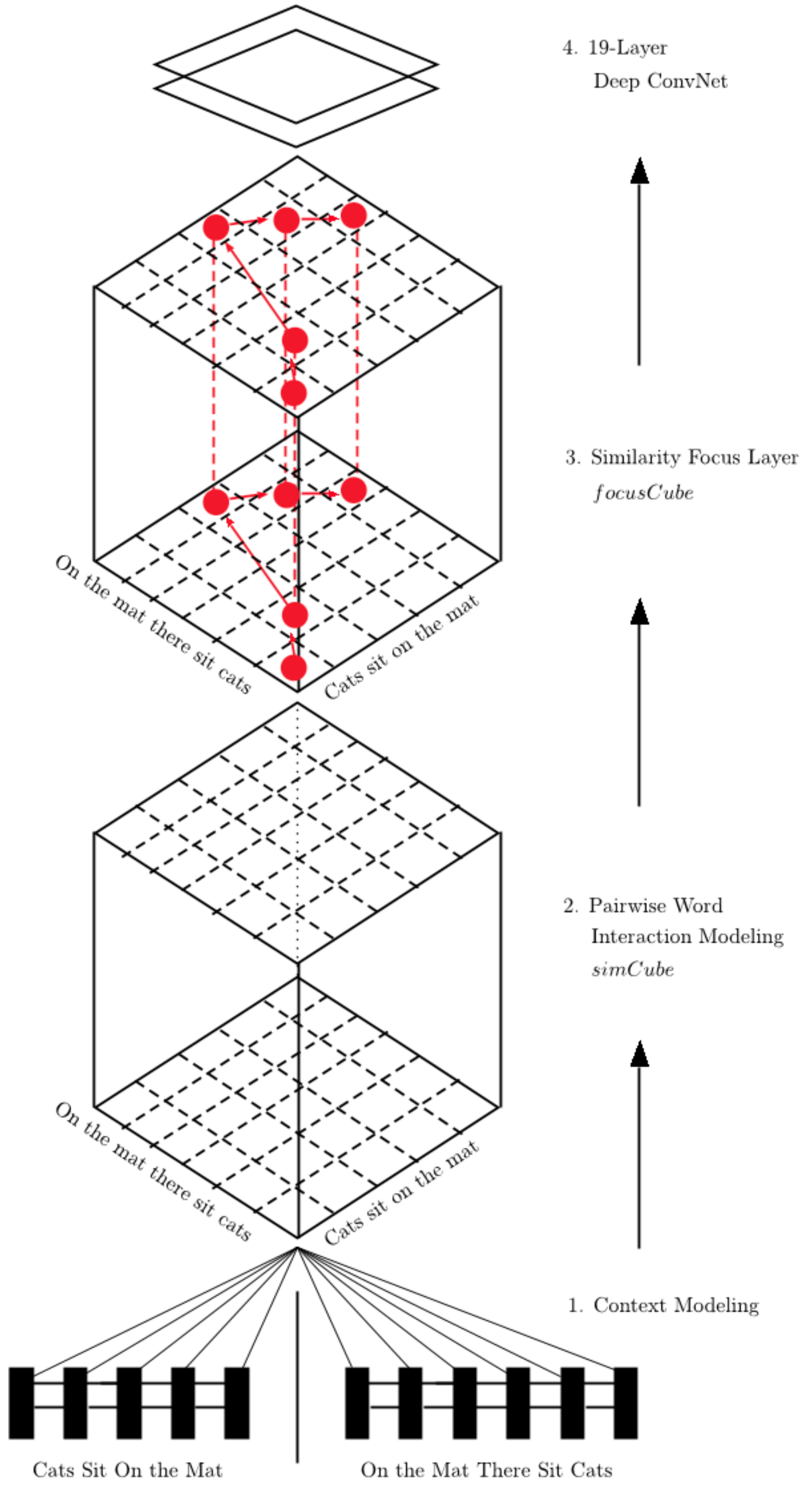
Deep Pairwise Word Model

- From Sentence Representation to Word Representation



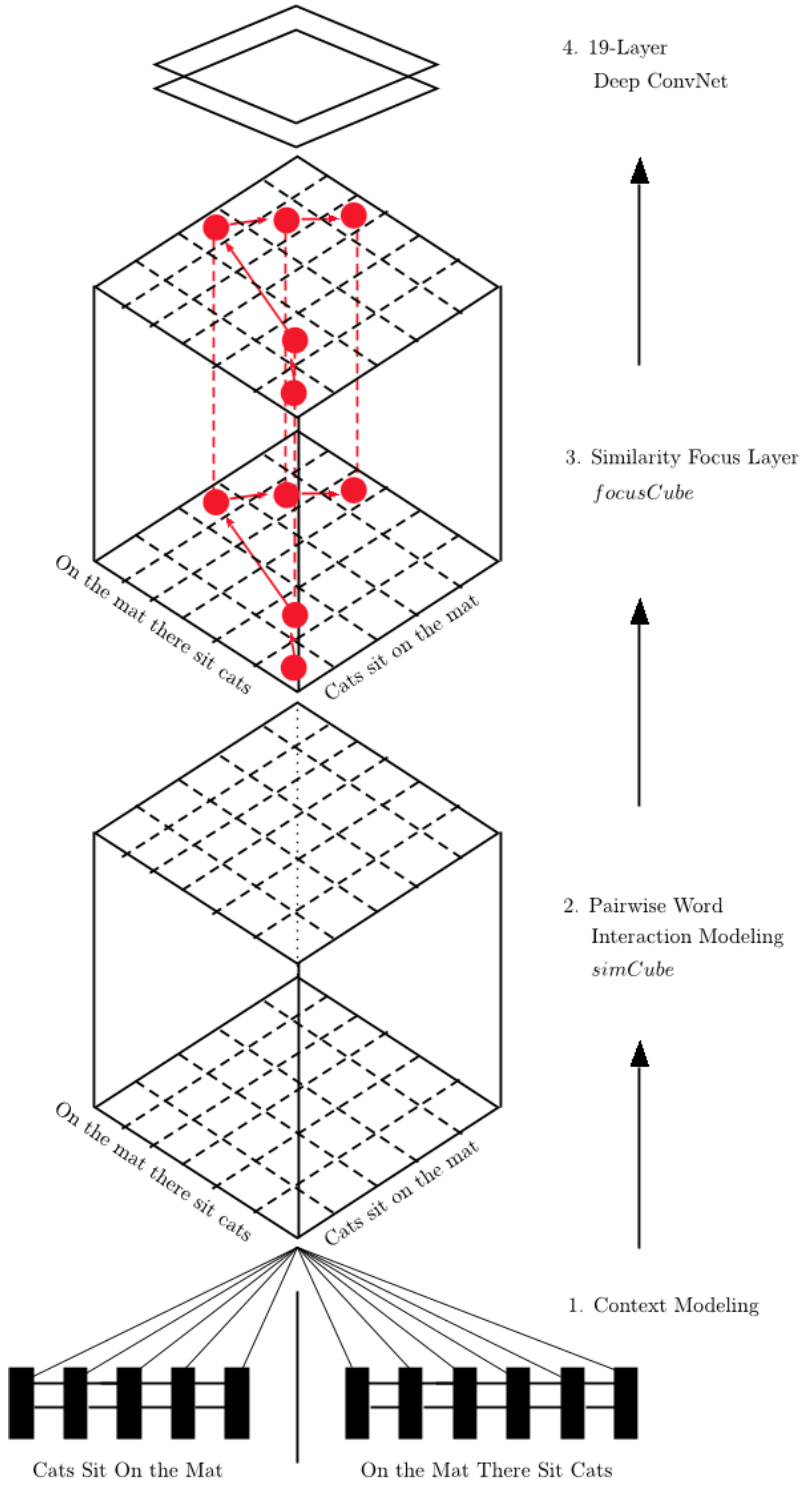
Deep Pairwise Word Model

- From Sentence Representation to Word Representation
- From Word Representation to Word Pair Interaction



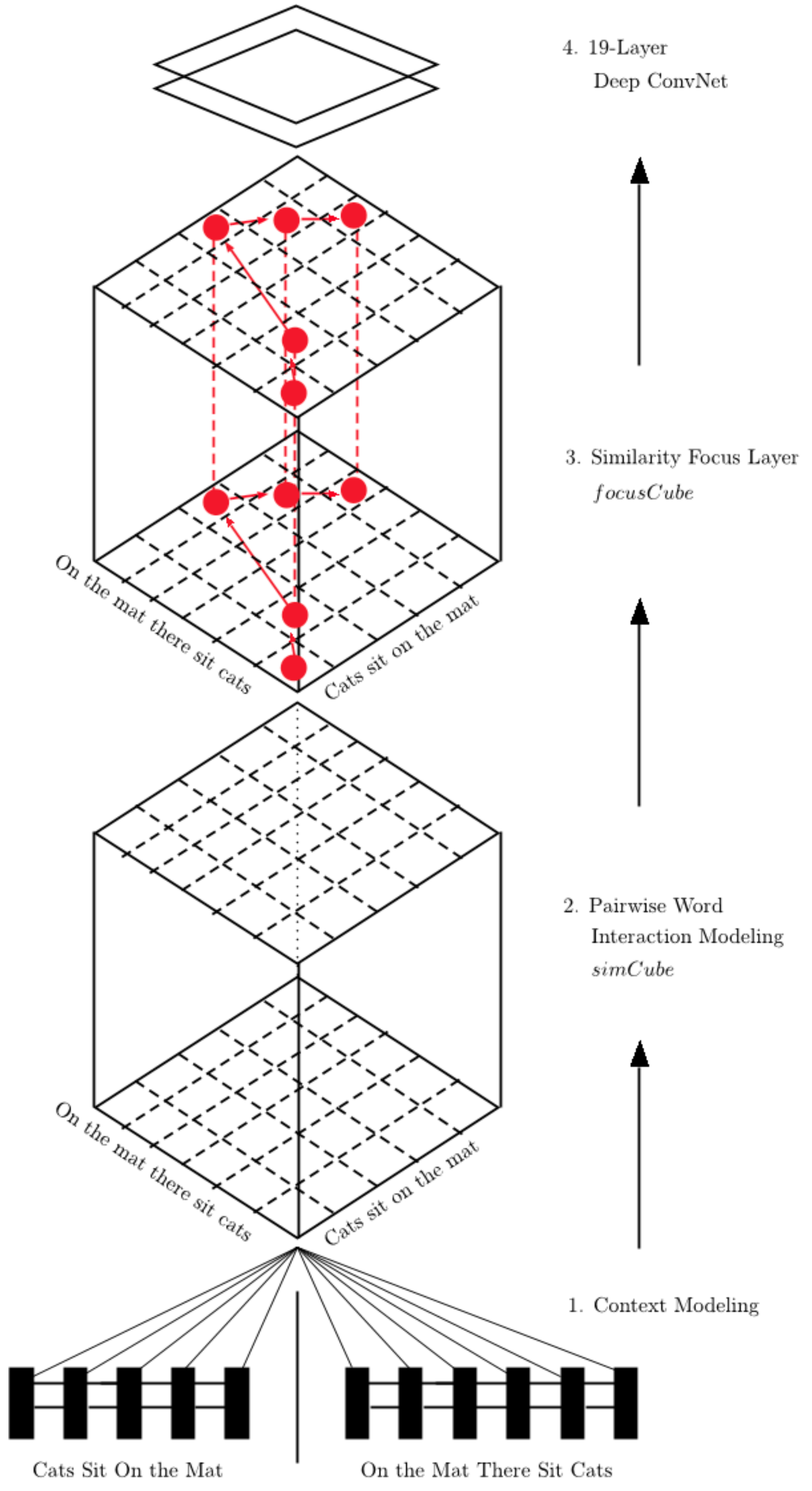
Deep Pairwise Word Model

- From Sentence Representation to Word Representation
- From Word Representation to Word Pair Interaction
- From Normal Interaction to Attentive Interaction

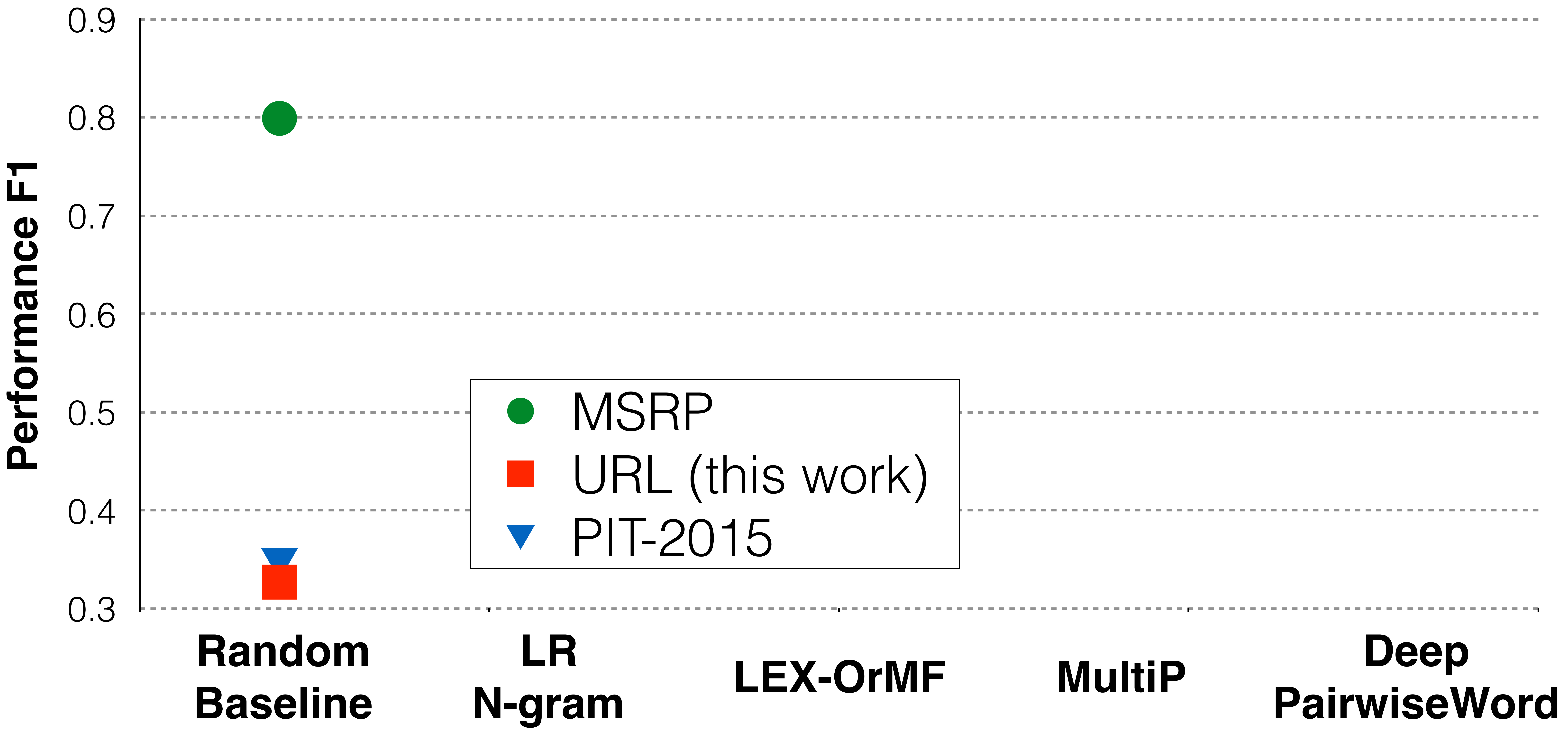


Deep Pairwise Word Model

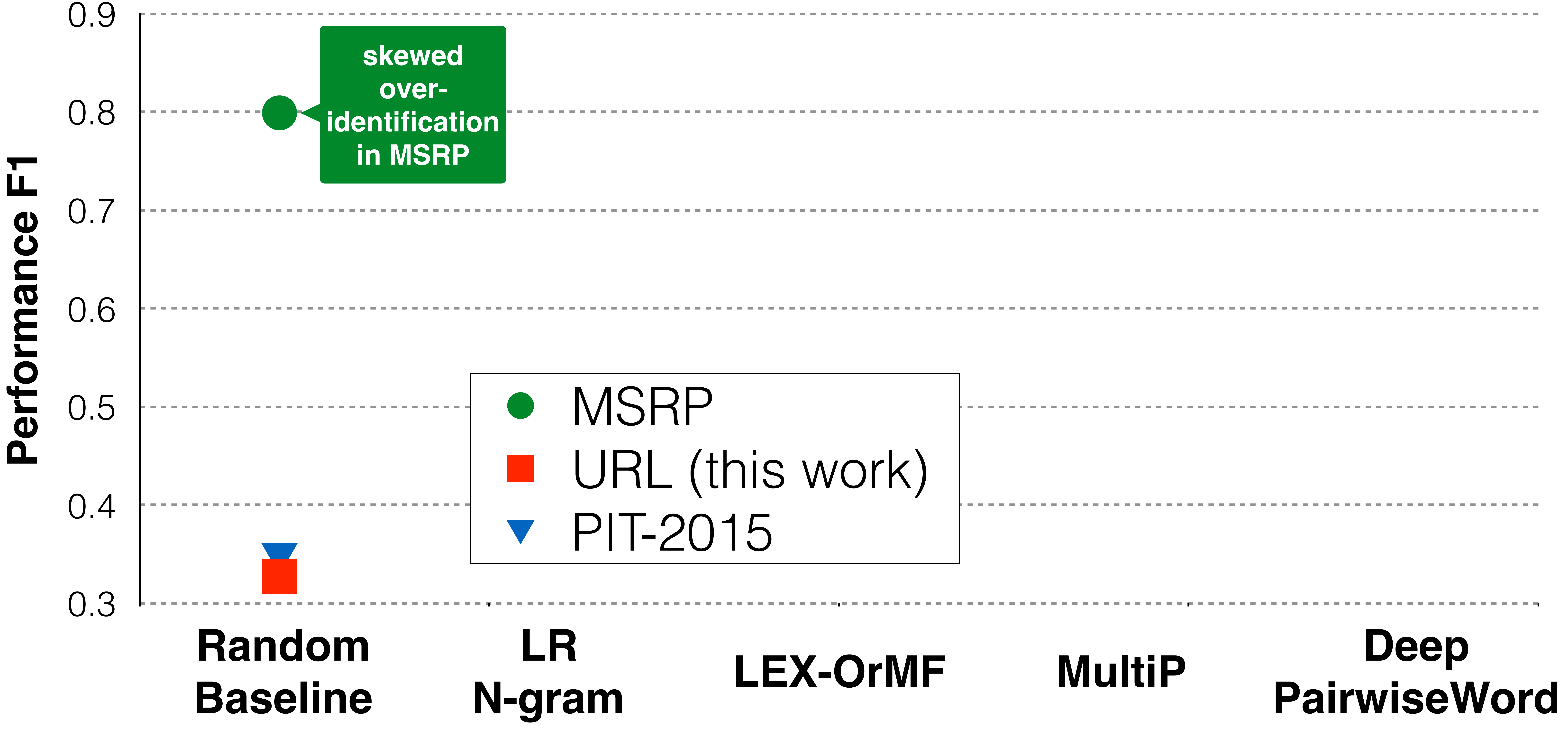
- From Sentence Representation to Word Representation
- From Word Representation to Word Pair Interaction
- From Normal Interaction to Attentive Interaction
- From Interaction to Pattern Recognition



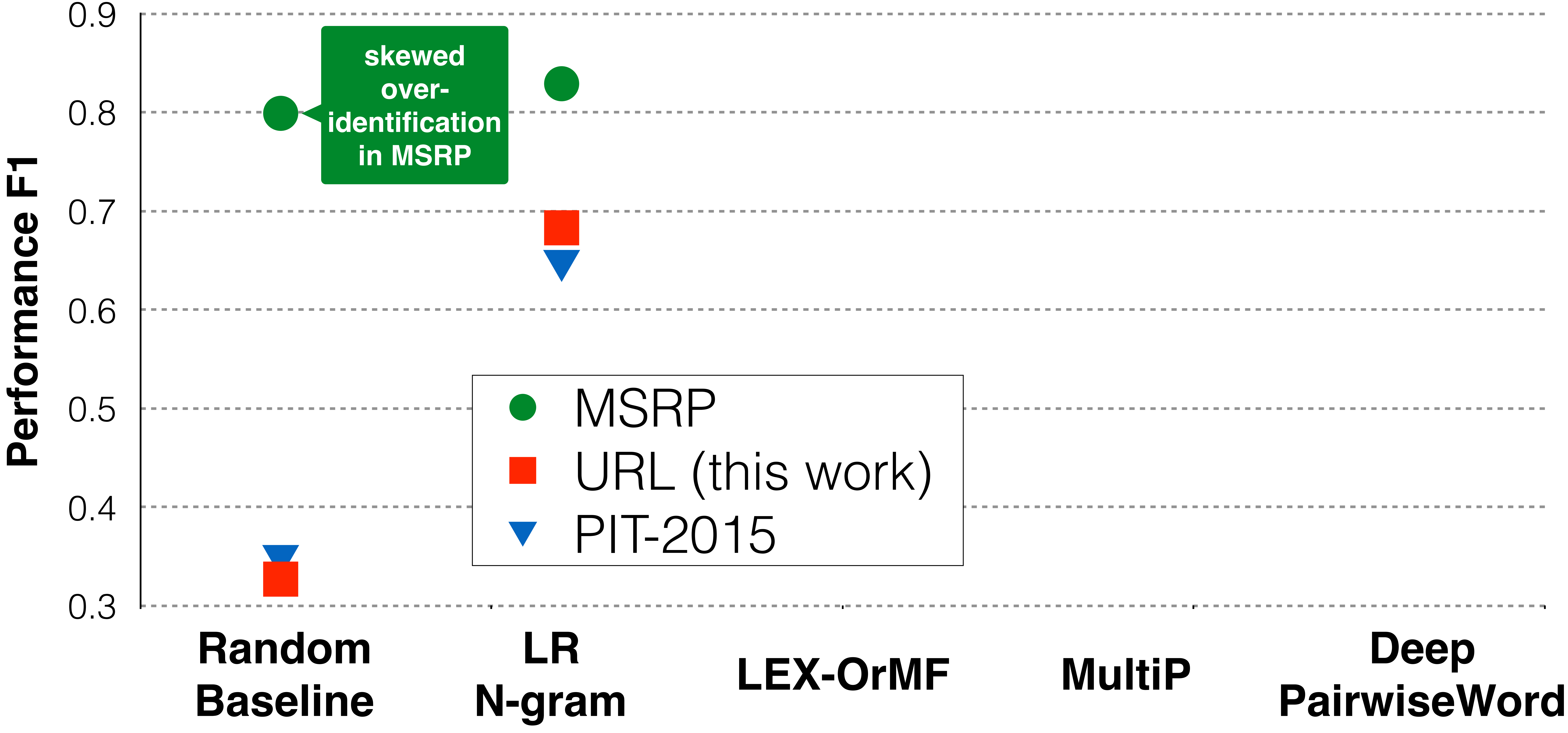
Automatic Paraphrase Identification



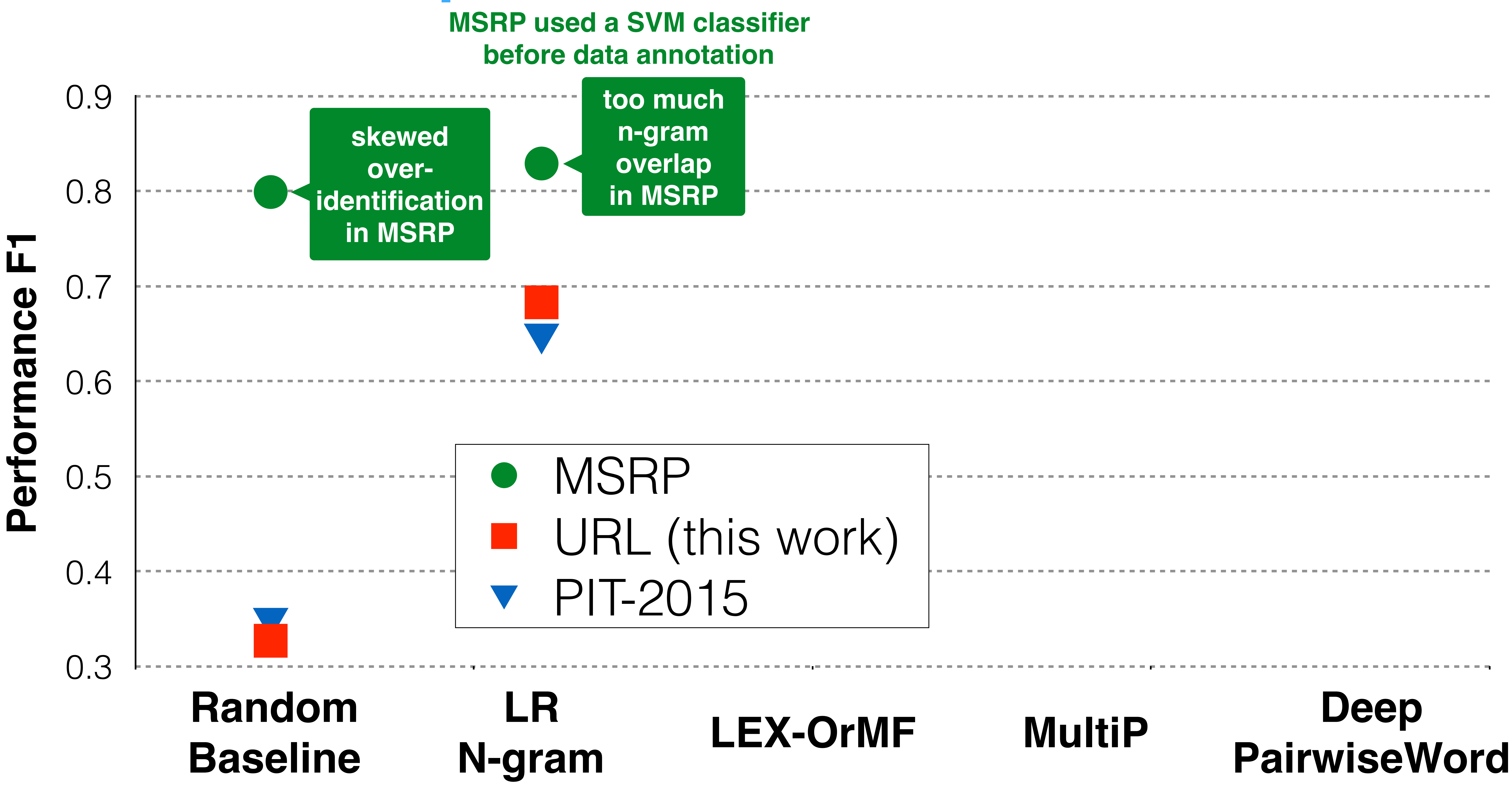
Automatic Paraphrase Identification



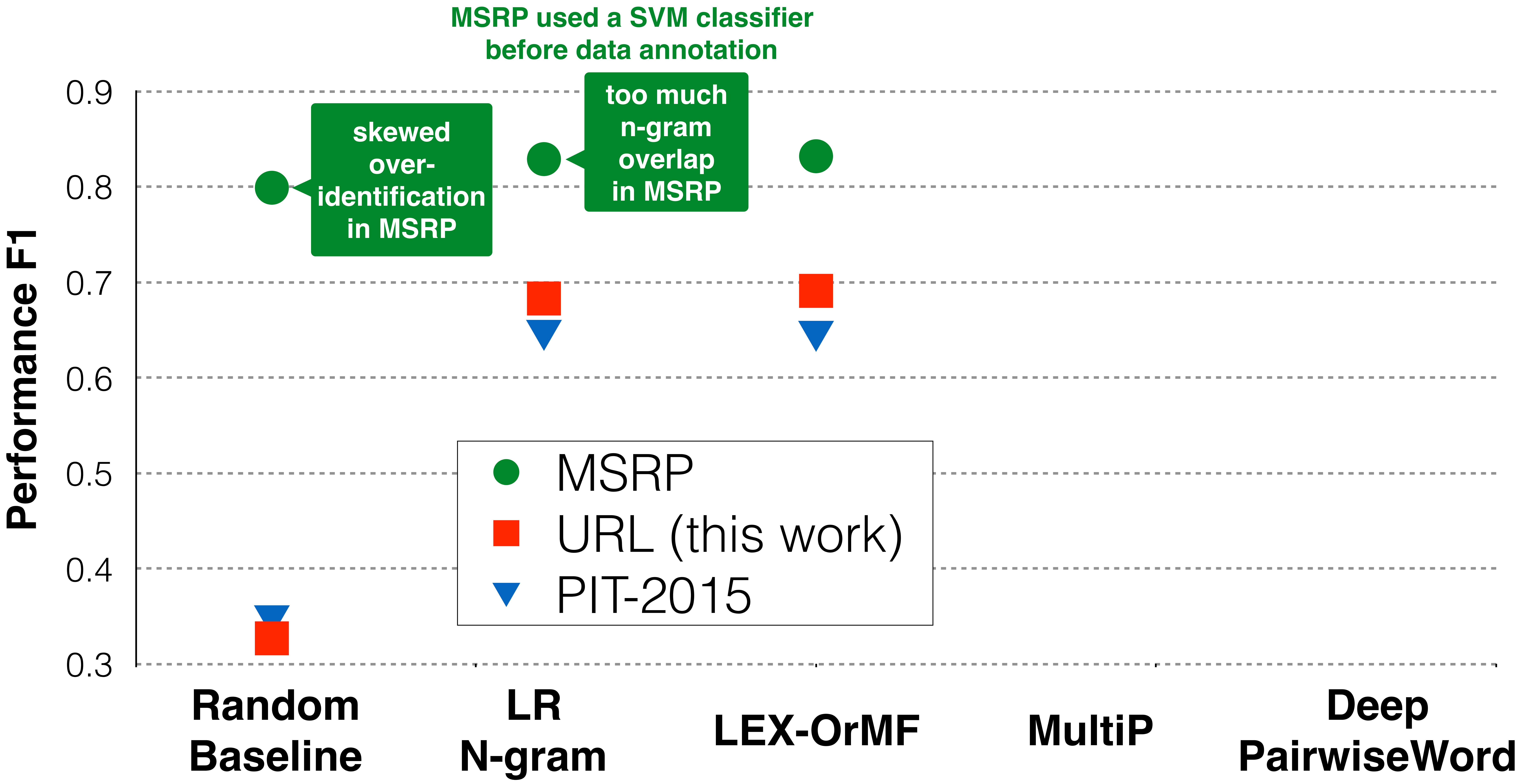
Automatic Paraphrase Identification



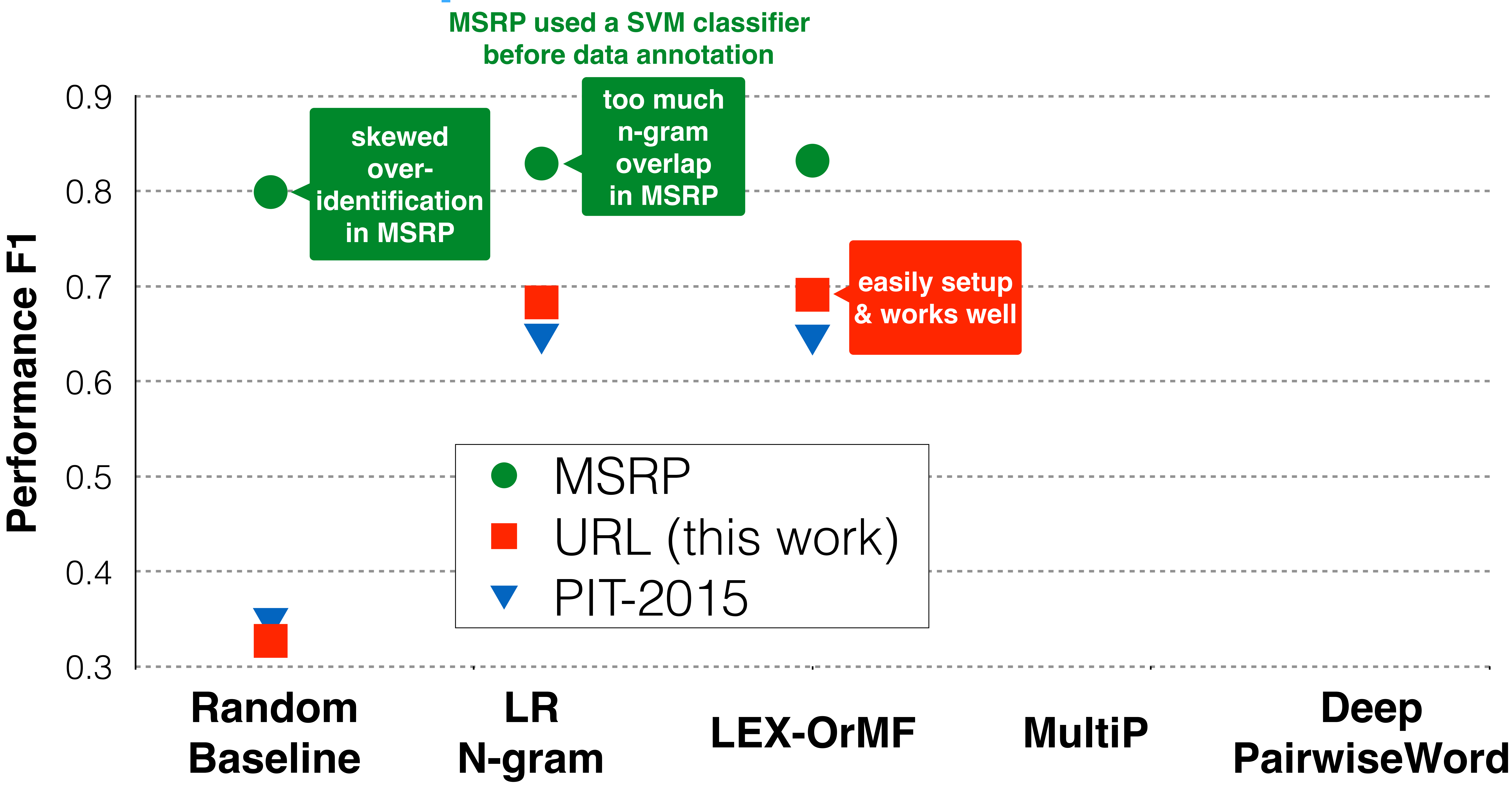
Automatic Paraphrase Identification



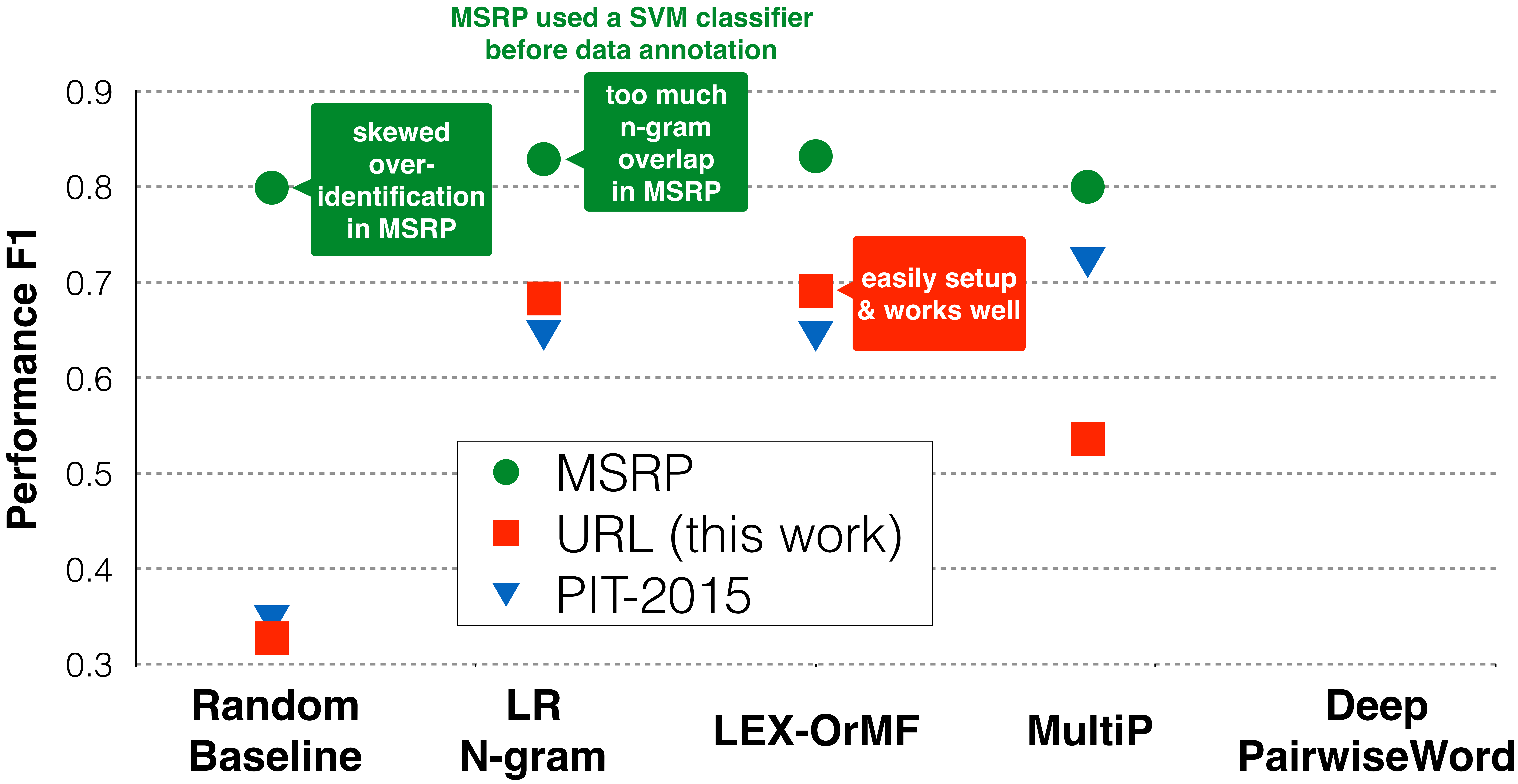
Automatic Paraphrase Identification



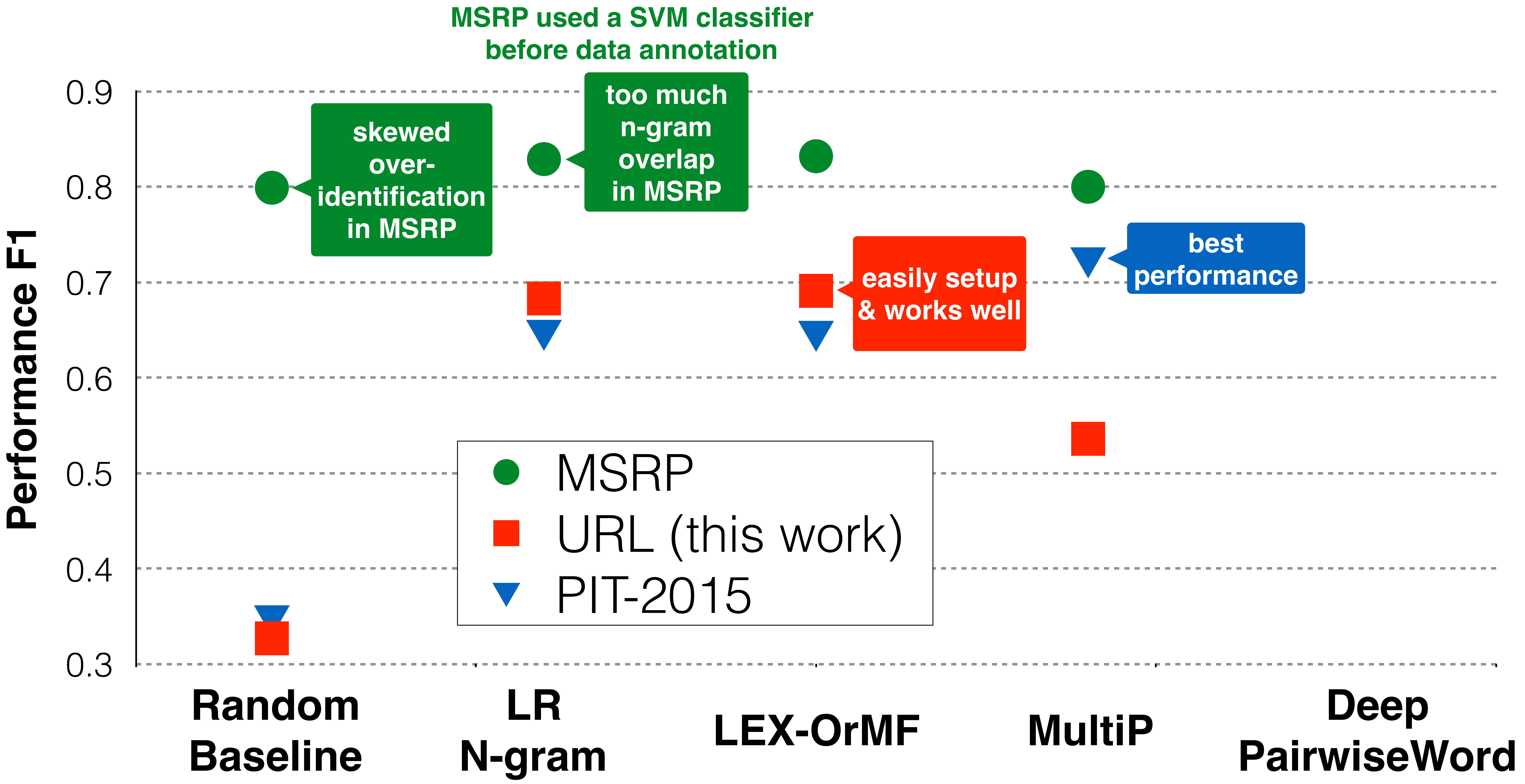
Automatic Paraphrase Identification



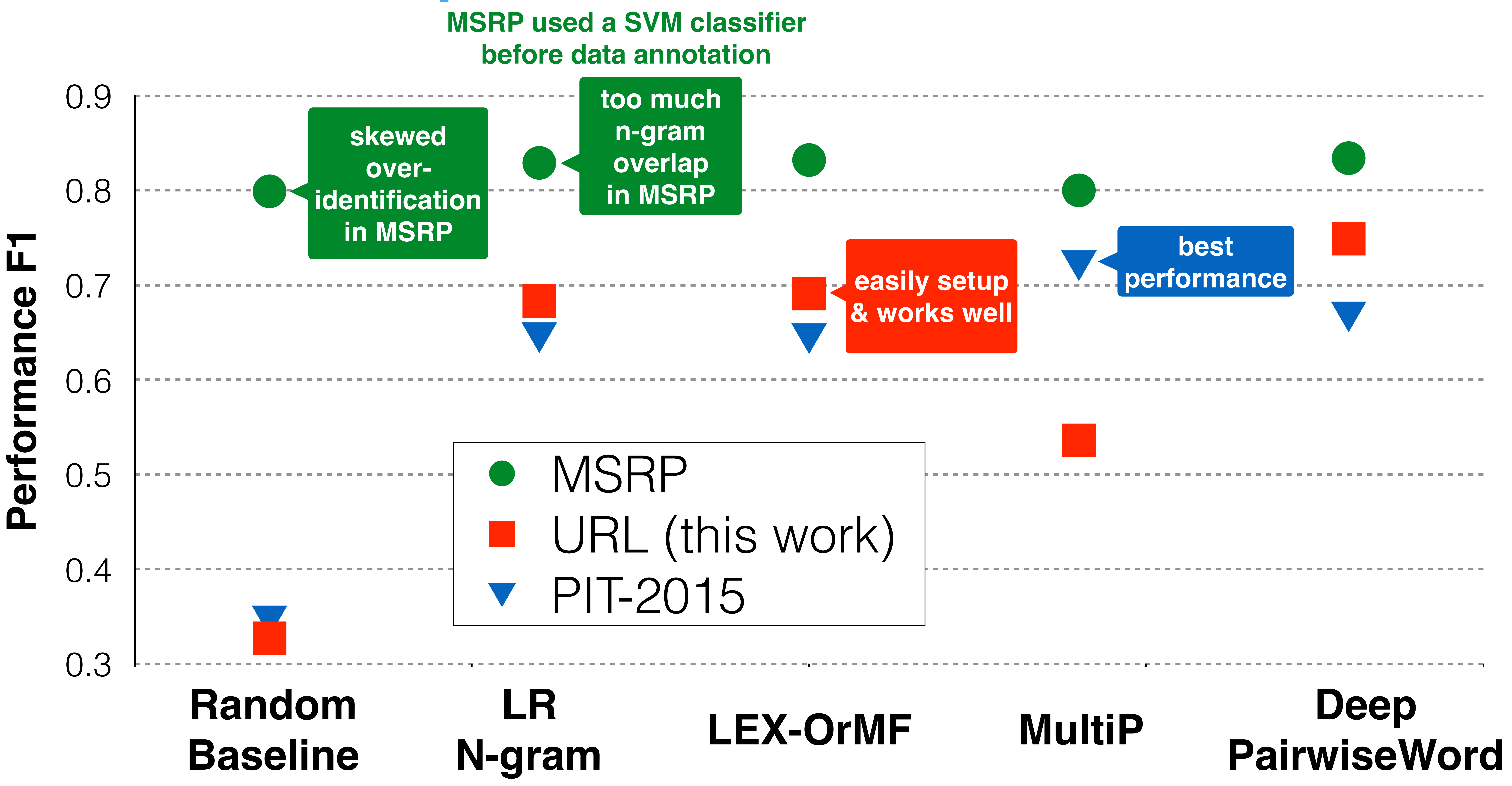
Automatic Paraphrase Identification



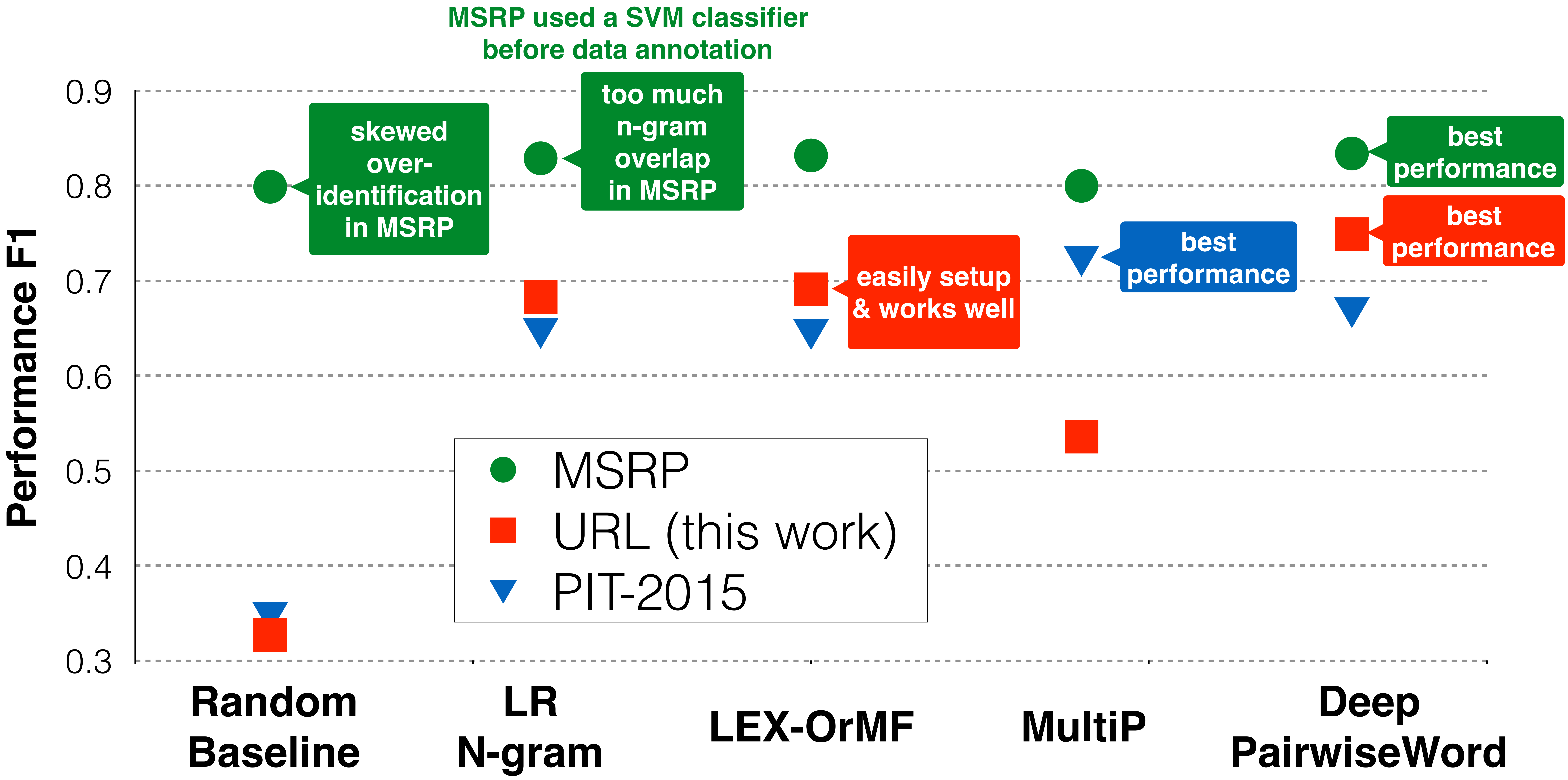
Automatic Paraphrase Identification



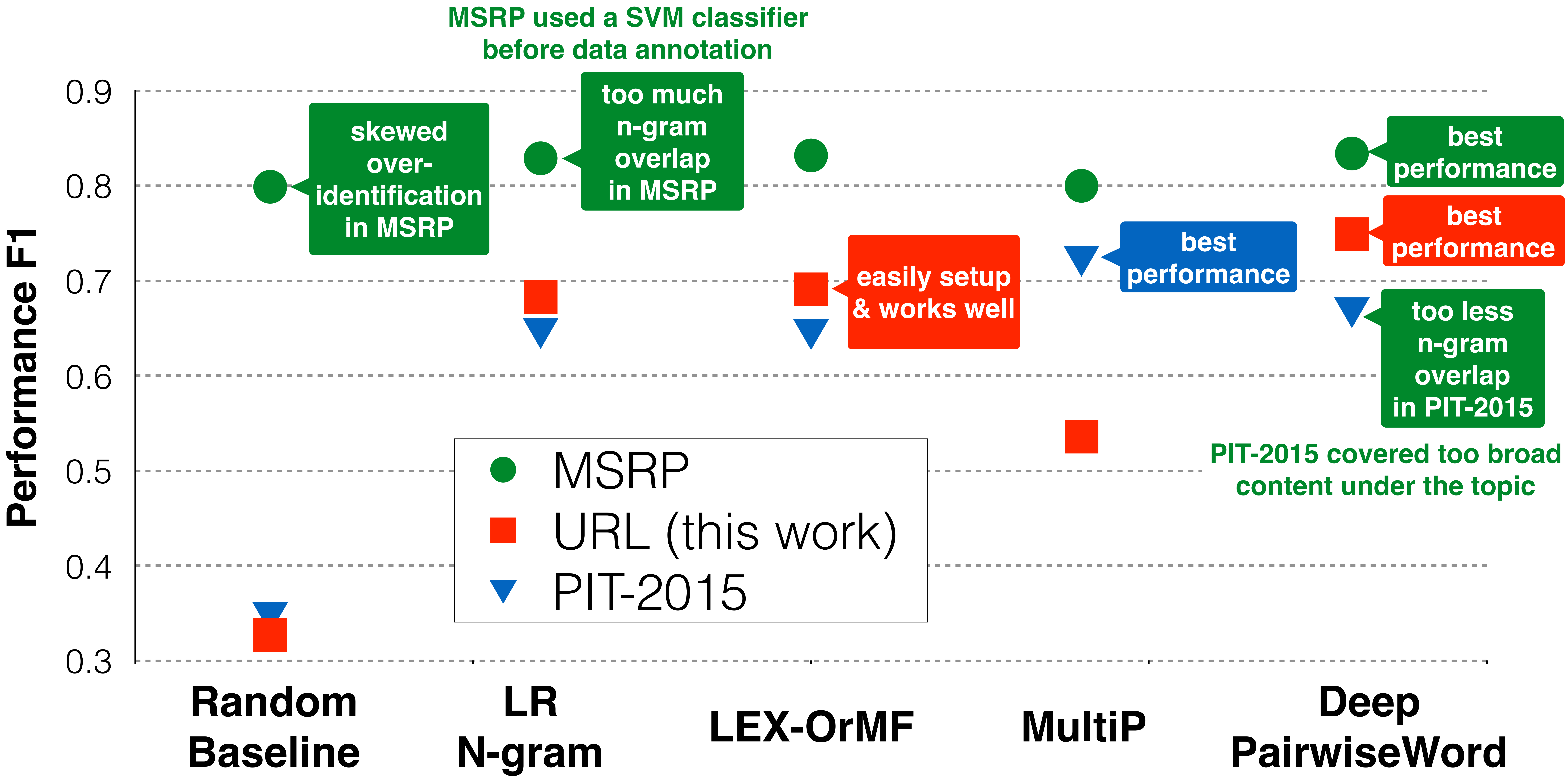
Automatic Paraphrase Identification



Automatic Paraphrase Identification

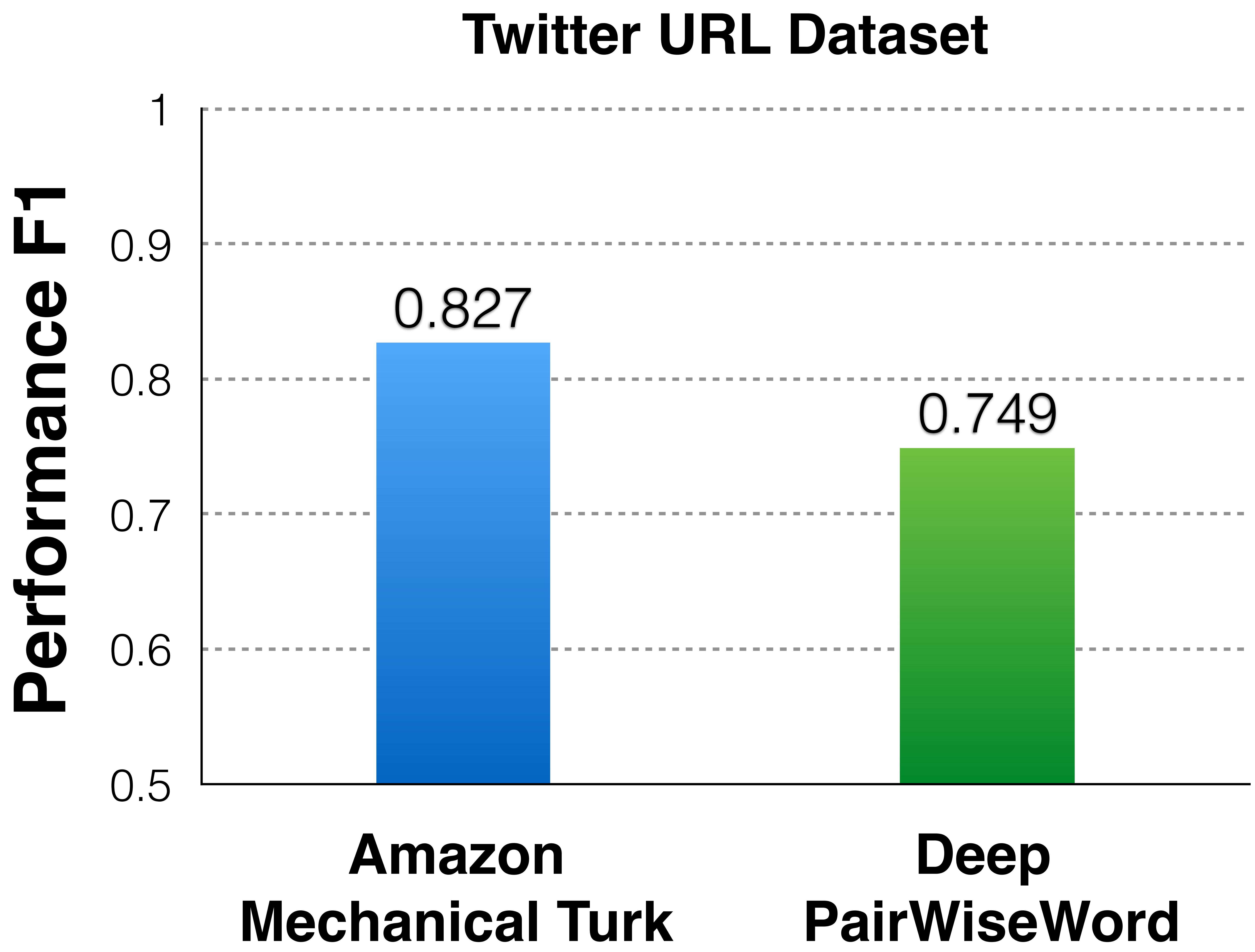


Automatic Paraphrase Identification



System Performance v.s. Human Upper-bound

System Performance v.s. Human Upper-bound



Error Analysis: Falsely Negative

This newly discovered species of moth has been named after Donald Trump.

New #moth named in honor of Donald Trump @realDonaldTrump

Error Analysis: Falsely Negative

This newly discovered species of moth has been named after Donald Trump.

New #moth named in honor of Donald Trump @realDonaldTrump

Error Analysis: Falsely Negative

This newly discovered species of moth has been named after Donald Trump.

New #moth named in honor of Donald Trump @realDonaldTrump

Out-of-Vocabulary Word Problem

Out-of-Vocabulary Word Problem

Dataset	Training Size	Test Size	# INV	# OOV	OOV Ratio	Source
PIT-2015	11530	838	7771	1238	13.7%	Twitter trends
Twitter-URL	42200	9324	24905	11440	31.5%	Twitter/news
MSRP	4076	1725	16226	1614	9.0%	news

Out-of-Vocabulary Word Problem

Dataset	Training Size	Test Size	# INV	# OOV	OOV Ratio	Source
PIT-2015	11530	838	7771	1238	13.7%	Twitter trends
Twitter-URL	42200	9324	24905	11440	31.5%	Twitter/news
MSRP	4076	1725	16226	1614	9.0%	news

Out-of-Vocabulary Word Problem

Dataset	Training Size	Test Size	# INV	# OOV	OOV Ratio	Source
PIT-2015	11530	838	7771	1238	13.7%	Twitter trends
Twitter-URL	42200	9324	24905	11440	31.5%	Twitter/news
MSRP	4076	1725	16226	1614	9.0%	news

Randomly initialized word embeddings fail to capture word syntax and semantics

Representing Word with Smaller Units

Representing Word with Smaller Units

Unit	**Output of σ(brexit)**
unigram	b, r, e, x, i, t
bigram w overlap	br, re, ex, xi, it
bigram w/o overlap	br, ex, it
trigram w overlap	bre, rex, exi, xit
trigram w/o overlap	bre, xit
whole word	brexit

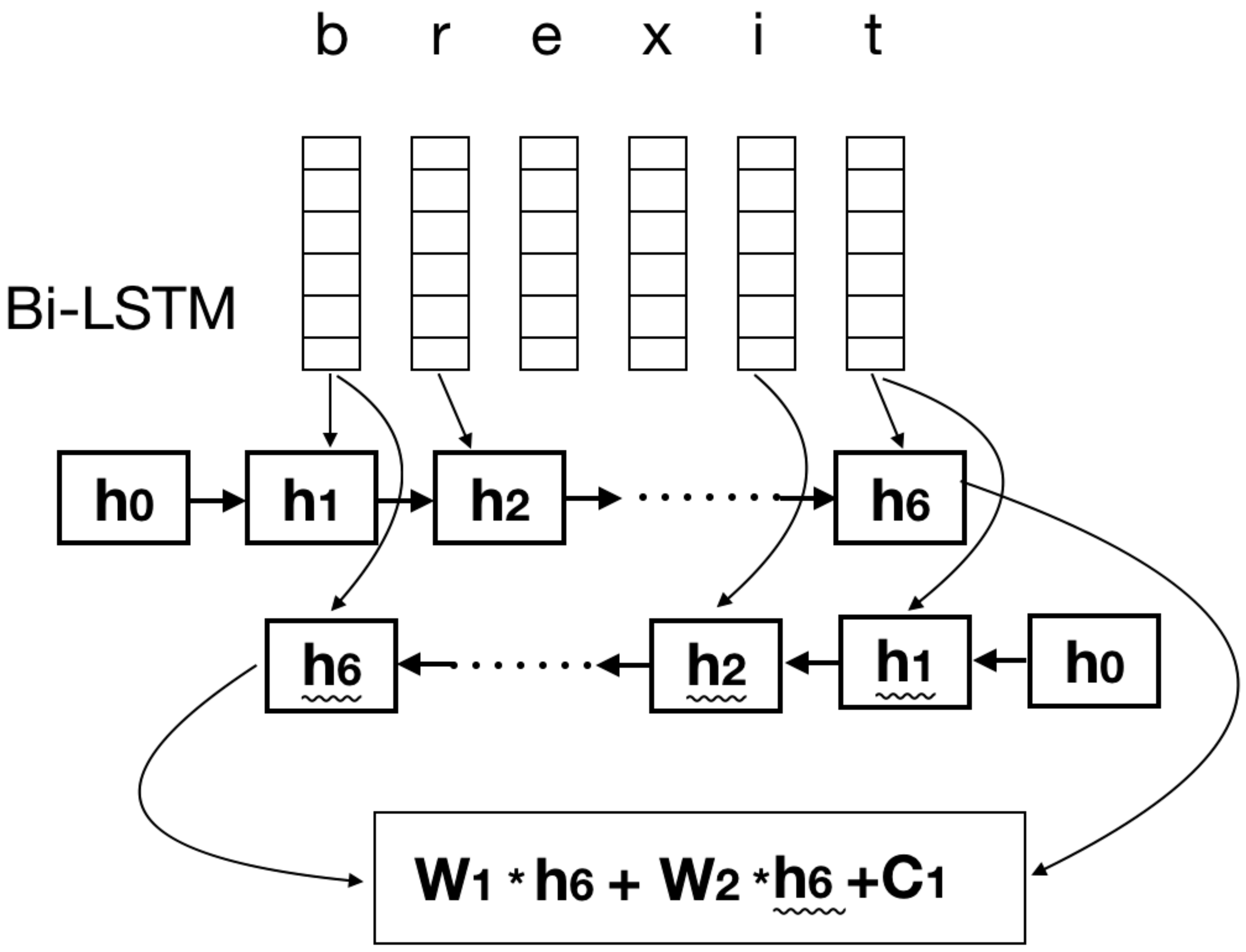
Table 1: Ngram examples for word ***brexit***.

LSTM Based Character Embedding (C2W)[1]

LSTM Based Character Embedding (C2W)[1]

[1] Ling et al., 2015

LSTM Based Character Embedding (C2W)[1]



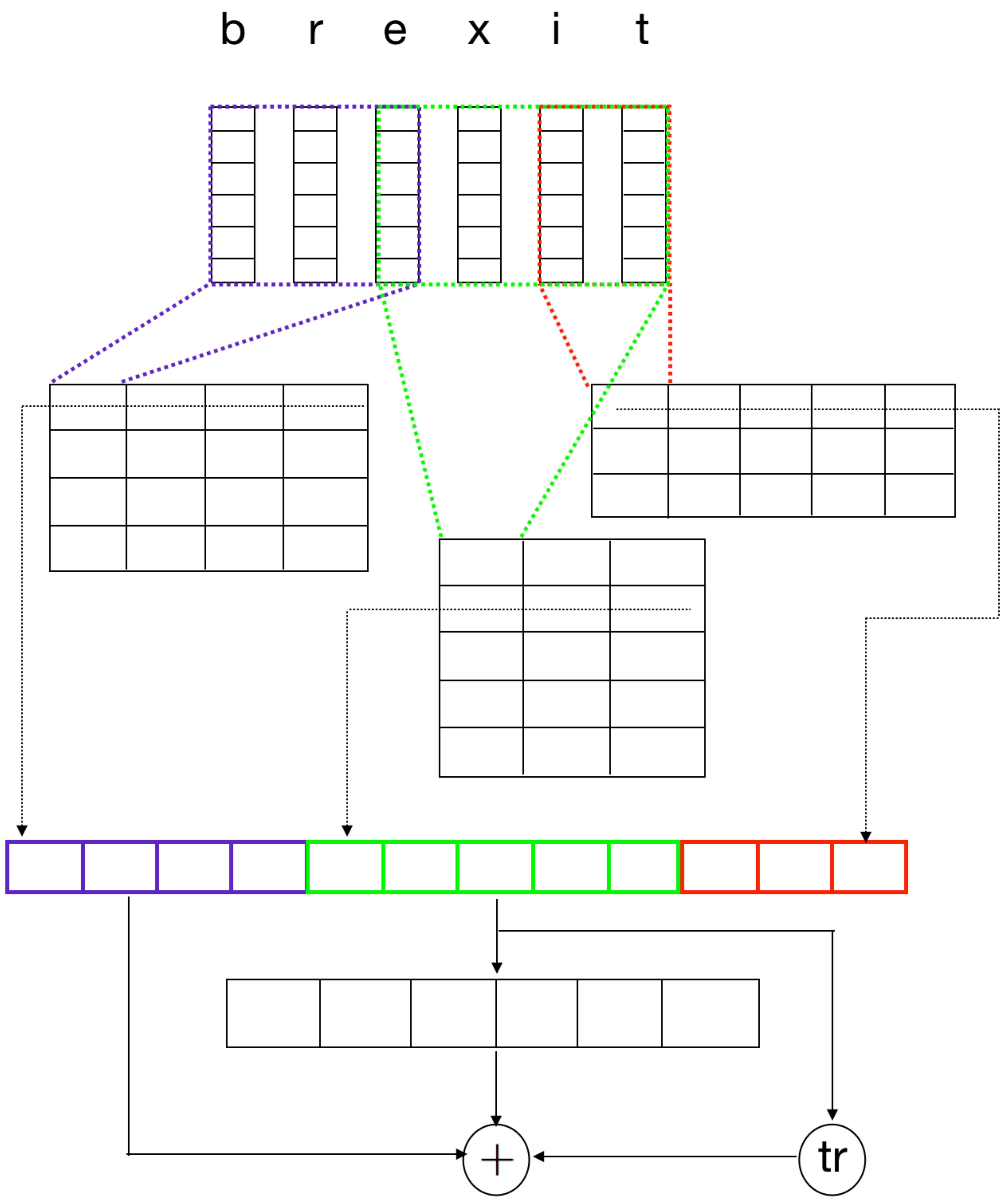
[1] Ling et al., 2015

CNN Based Character Embedding[1]

CNN Based Character Embedding[1]

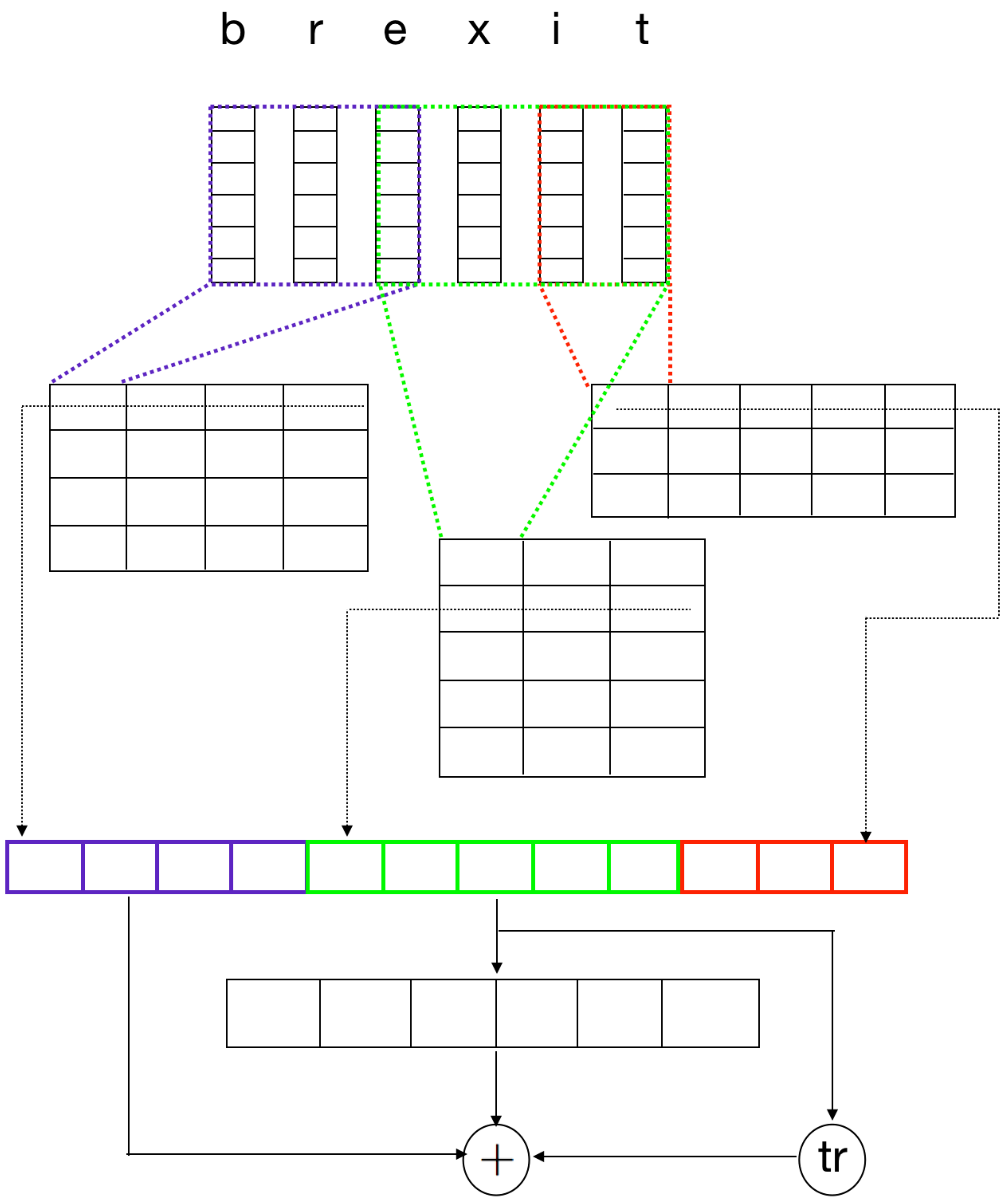
[1] Kim et al., 2016

CNN Based Character Embedding[1]



[1] Kim et al., 2016

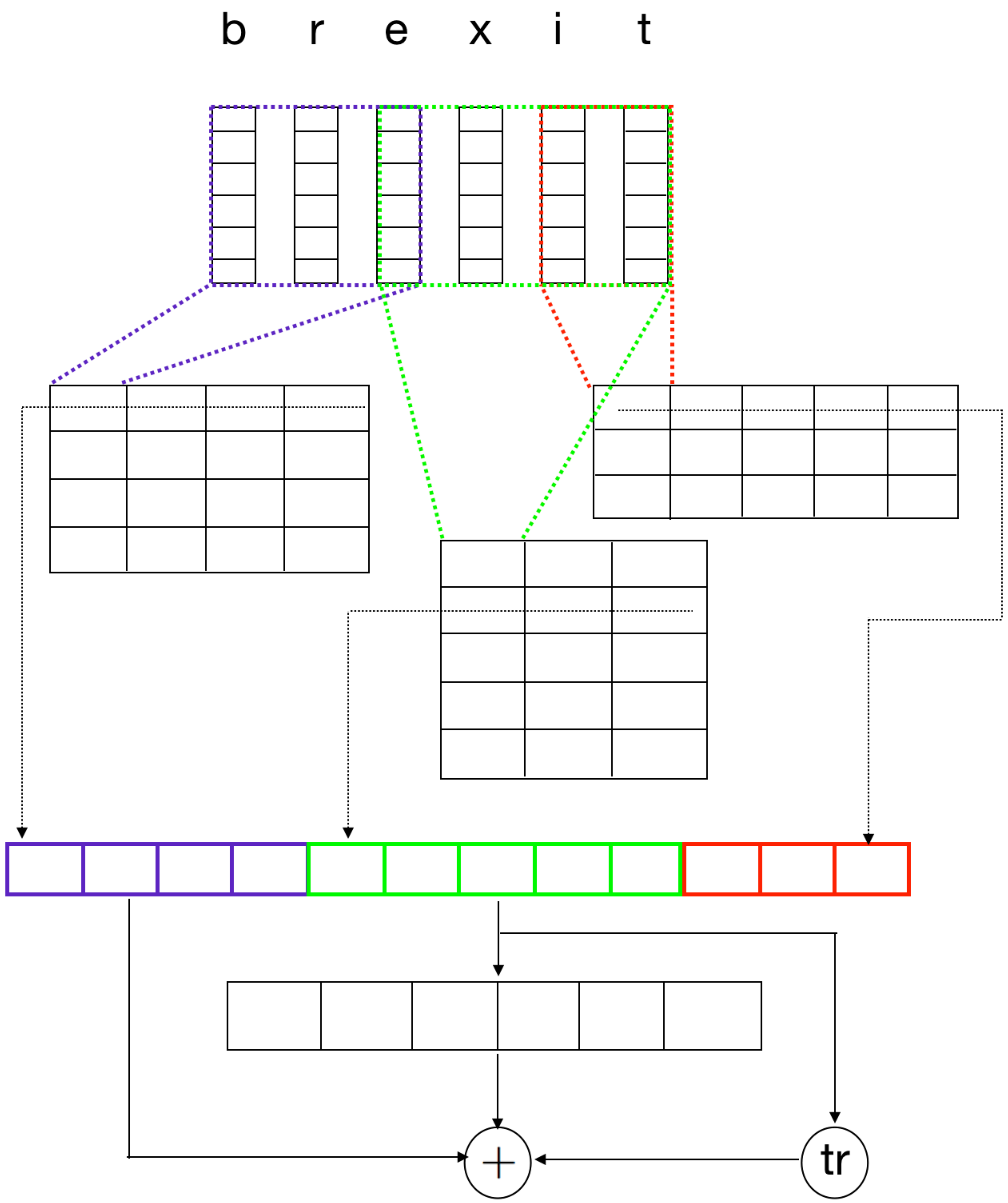
CNN Based Character Embedding[1]



Embedding Concatenation

[1] Kim et al., 2016

CNN Based Character Embedding[1]



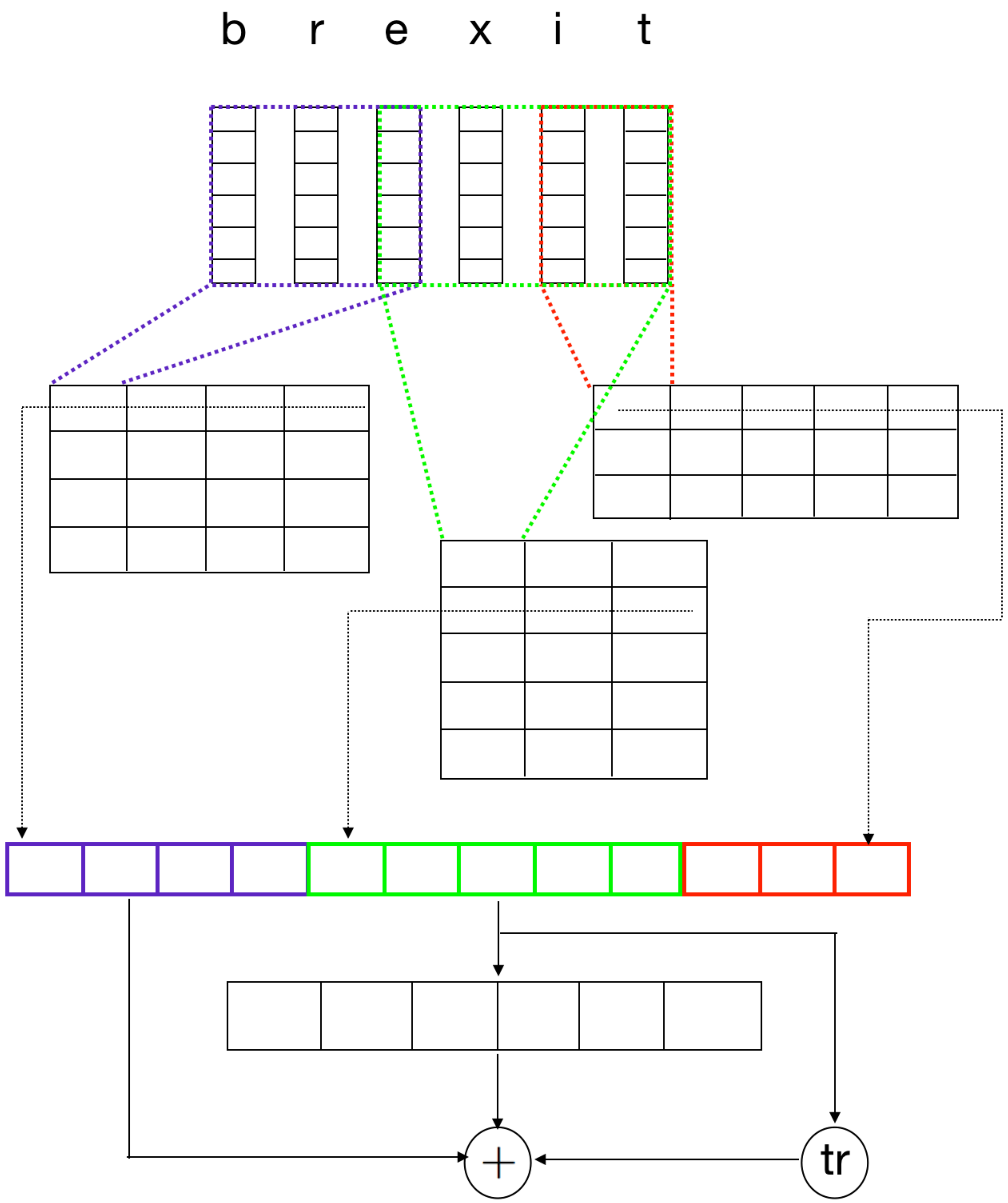
Embedding Concatenation

Convolution with multiple filters

$$\mathbf{f}^k[i] = \tanh(\langle \mathbf{C}^k[*, i : i + w - 1], \mathbf{H} \rangle + b)$$

[1] Kim et al., 2016

CNN Based Character Embedding[1]



Embedding Concatenation

Convolution with multiple filters

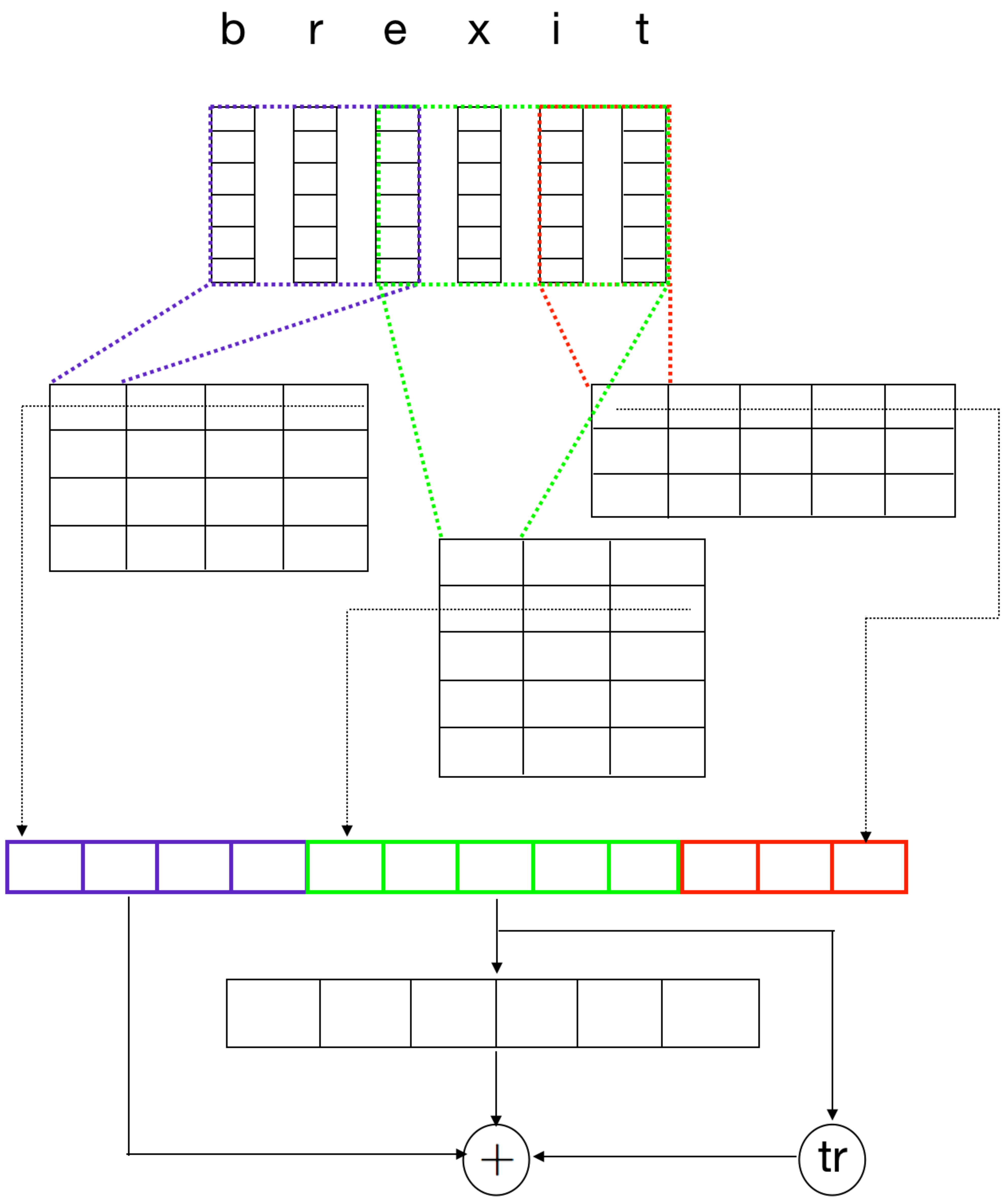
$$\mathbf{f}^k[i] = \tanh(\langle \mathbf{C}^k[*, i : i + w - 1], \mathbf{H} \rangle + b)$$

max pooling

$$y^k = \max_i \mathbf{f}^k[i]$$

[1] Kim et al., 2016

CNN Based Character Embedding[1]



Embedding Concatenation

Convolution with multiple filters

$$\mathbf{f}^k[i] = \tanh(\langle \mathbf{C}^k[*, i : i + w - 1], \mathbf{H} \rangle + b)$$

max pooling

$$y^k = \max_i \mathbf{f}^k[i]$$

highway network

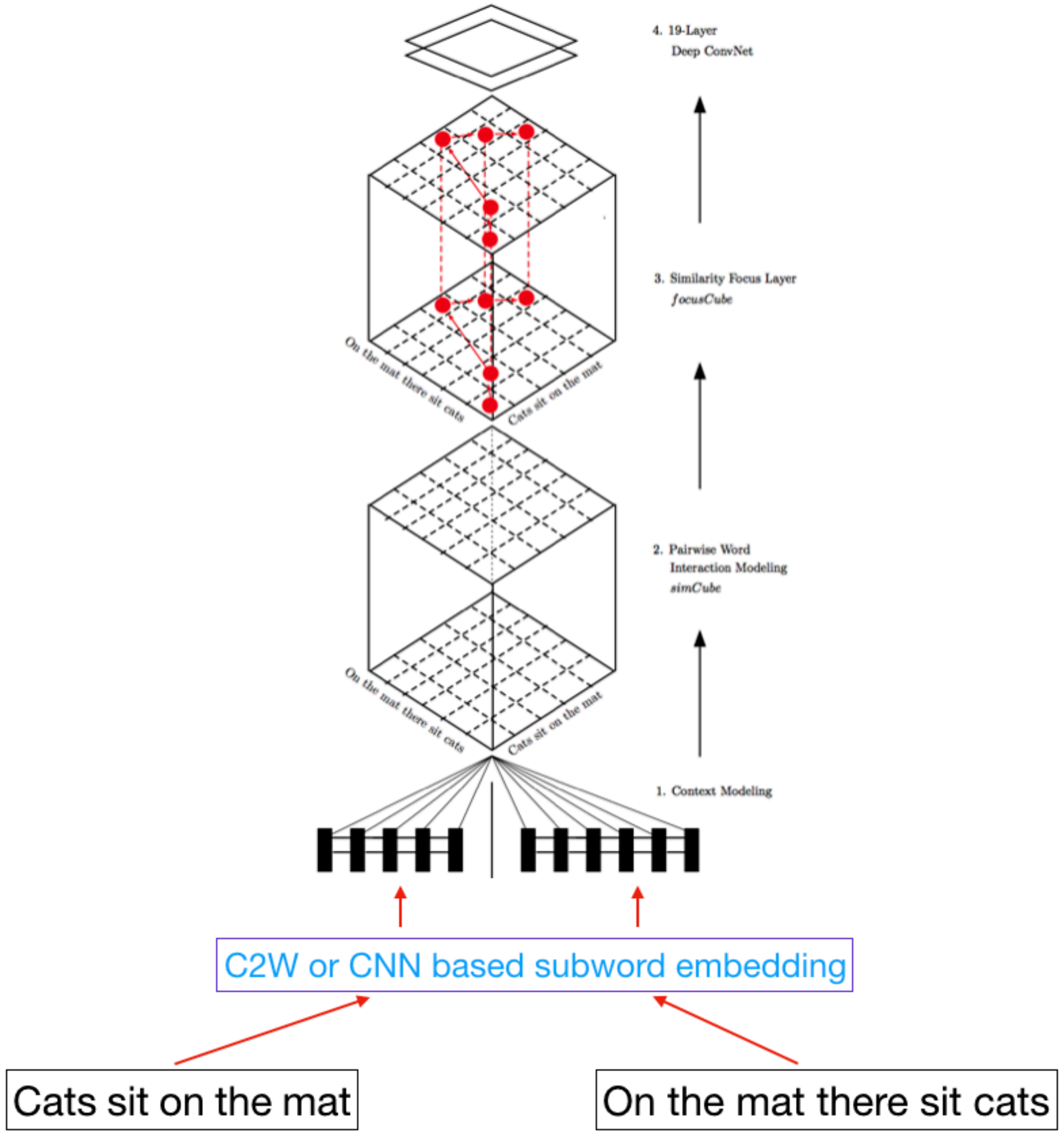
$$\mathbf{t} = \sigma(\mathbf{W}_T \mathbf{y} + \mathbf{b}_T)$$

$$\mathbf{z} = \mathbf{t} \odot g(\mathbf{W}_H \mathbf{y} + \mathbf{b}_H) + (\mathbf{1} - \mathbf{t}) \odot \mathbf{y}$$

[1] Kim et al., 2016

Subword Based Pairwise Word Interaction Model

Subword Based Pairwise Word Interaction Model



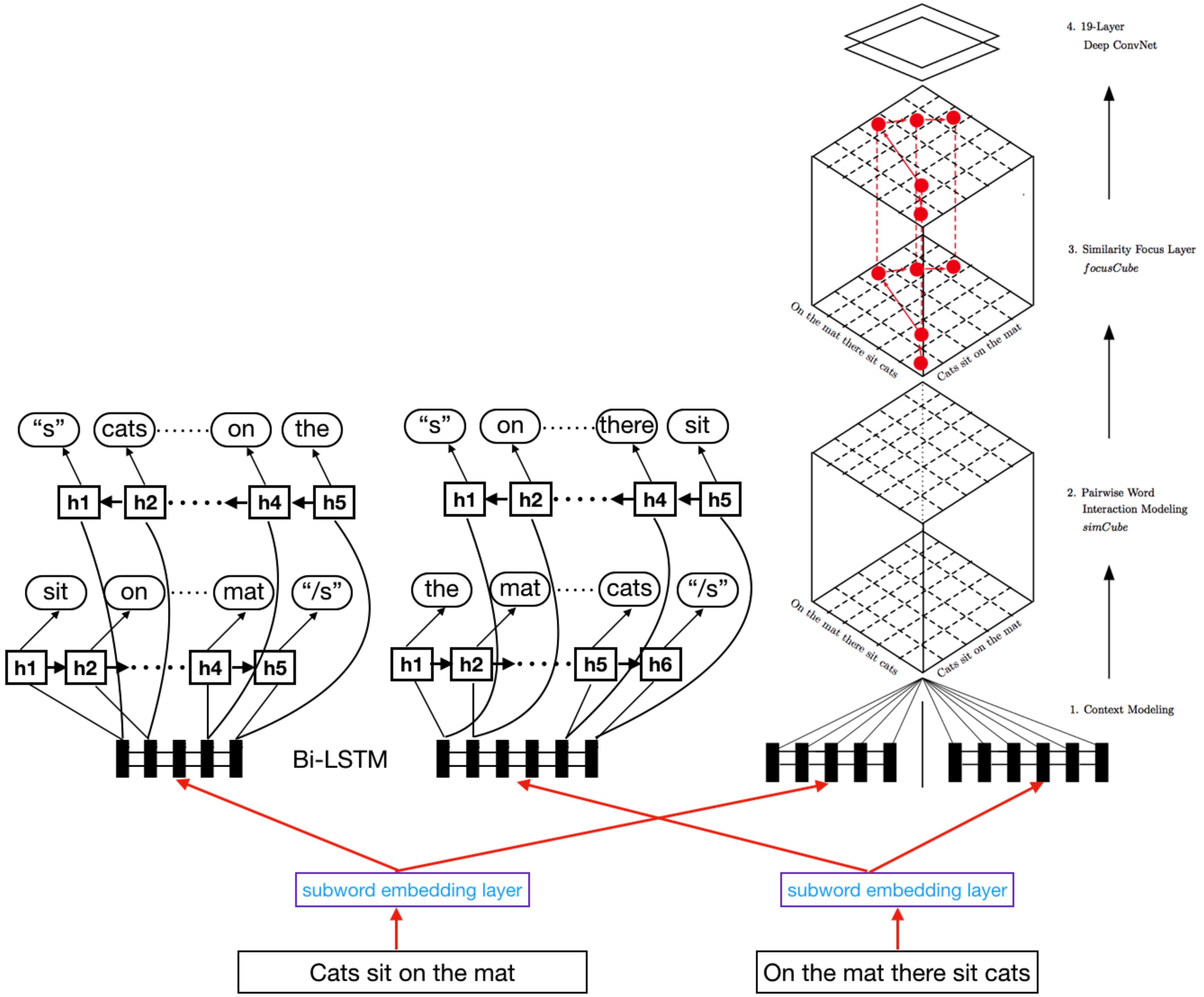
Word Embedding v.s. Subword Embedding

Word Embedding v.s. Subword Embedding

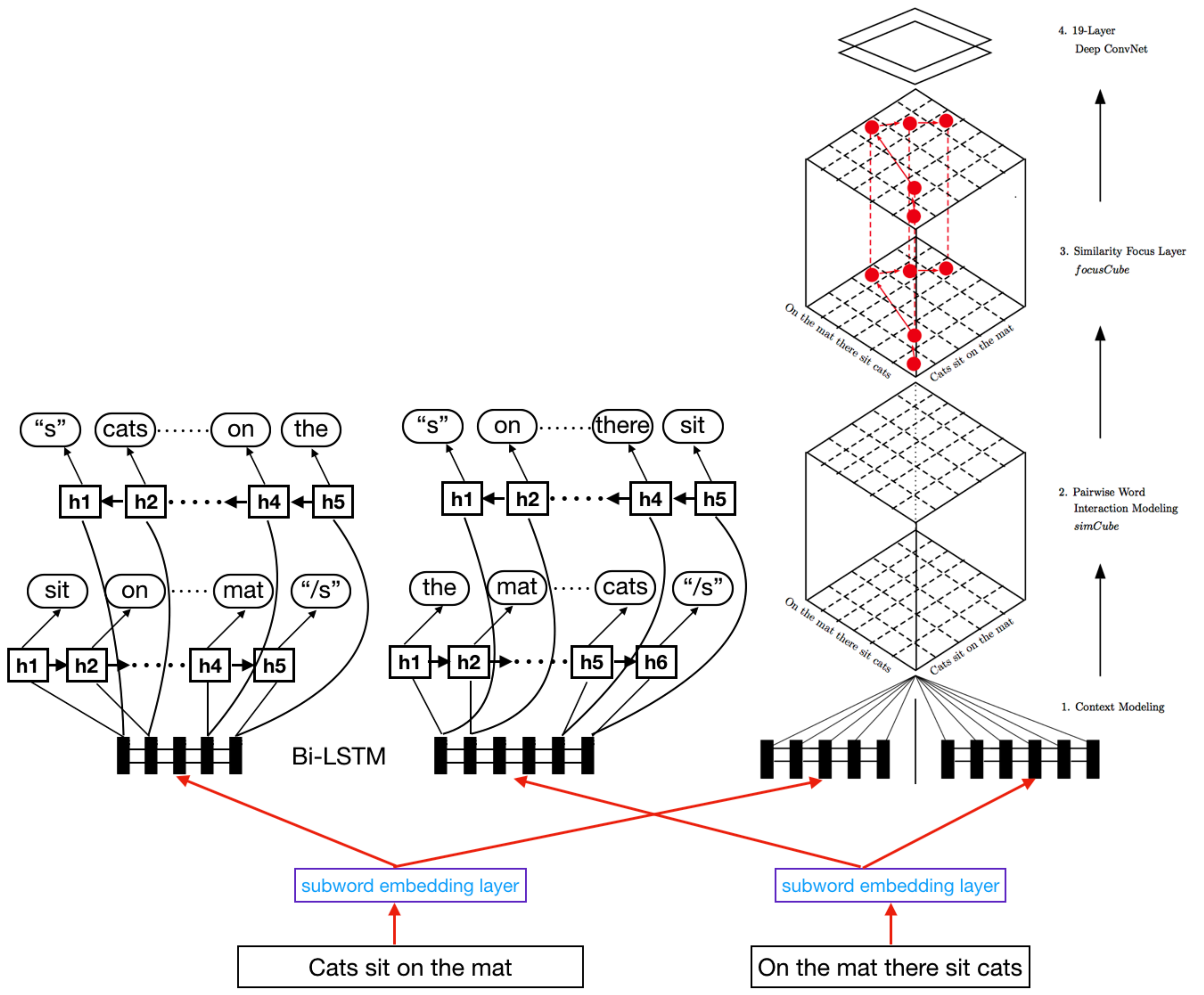
	Model Variations	pre-train	#parameters	Twitter URL	PIT-2015	MSRP
Word Models	Logistic Regression	-	-	0.683	0.645	0.829
	(Lan et al., 2017)	Yes	9.5M	0.749	0.667	0.834
	pretrained, fixed	Yes	2.2M	0.753	0.632	0.834
	pretrained, updated	Yes	9.5M	0.756	0.656	0.832
	randomized, fixed	–	2.2M	0.728	0.456	0.821
	randomized, updated	–	9.5M	0.735	0.625	0.834
Subword Models	C2W, unigram	–	2.6M	0.742	0.534	0.816
	C2W, bigram	–	2.7M	0.742	0.563	0.825
	C2W, trigram	–	3.1M	0.729	0.576	0.824
	CNN, unigram	–	6.5M	0.756	0.589	0.820
	CNN, bigram	–	6.5M	0.760	0.646	0.814
	CNN, trigram	–	6.7M	0.753	0.667	0.818

Multi-task Language Model

Multi-task Language Model



Multi-task Language Model



$$E_{joint} = E + \gamma(\overrightarrow{E}_{LM} + \overleftarrow{E}_{LM})$$

New State-of-the-art with Multi-task Language Model

New State-of-the-art with Multi-task Language Model

	Model Variations	Pre-train	#Parameters	Twitter URL	PIT-2015	MSRP
Word Models	Logistic Regression	-	-	0.683	0.645	0.829
	(Lan et al., 2017)	Yes	9.5M	0.749	<u>0.667</u>	0.834
	pretrained, fixed	Yes	2.2M	0.753	0.632	0.834
	pretrained, updated	Yes	9.5M	0.756	0.656	0.832
	randomized, fixed	–	2.2M	0.728	0.456	0.821
	randomized, updated	–	9.5M	0.735	0.625	0.834
Subword Models	C2W, unigram	–	2.6M	0.742	0.534	0.816
	C2W, bigram	–	2.7M	0.742	0.563	0.825
	C2W, trigram	–	3.1M	0.729	0.576	0.824
	CNN, unigram	–	6.5M	0.756	0.589	0.820
	CNN, bigram	–	6.5M	0.760	0.646	0.814
	CNN, trigram	–	6.7M	0.753	<u>0.667</u>	0.818
Subword+LM	LM, C2W, unigram	–	3.5M	0.760	**0.691**	0.831
	LM, C2W, bigram	–	3.6M	**0.768**	0.651	0.830
	LM, C2W, trigram	–	4.0M	<u>0.765</u>	0.659	0.831
	LM, CNN, unigram	–	7.4M	0.754	0.665	**0.840**
	LM, CNN, bigram	–	7.4M	0.761	<u>0.667</u>	<u>0.835</u>
	LM, CNN, trigram	–	7.6M	0.759	<u>0.667</u>	0.831

Takeaways

- Simple but effective paraphrase collection method
- Largest annotated paraphrase corpora to date
- Continuously growing, providing up-to-date data
- Subword embedding for paraphrase identification
- Data and Code: https://github.com/lanwuwei/paraphrase-dataset


Q
&
A

Backup slides: Lexical Dissimilarity