

Social Media & Text Analysis

lecture 10 -

Convolutional Neural Networks and Attention

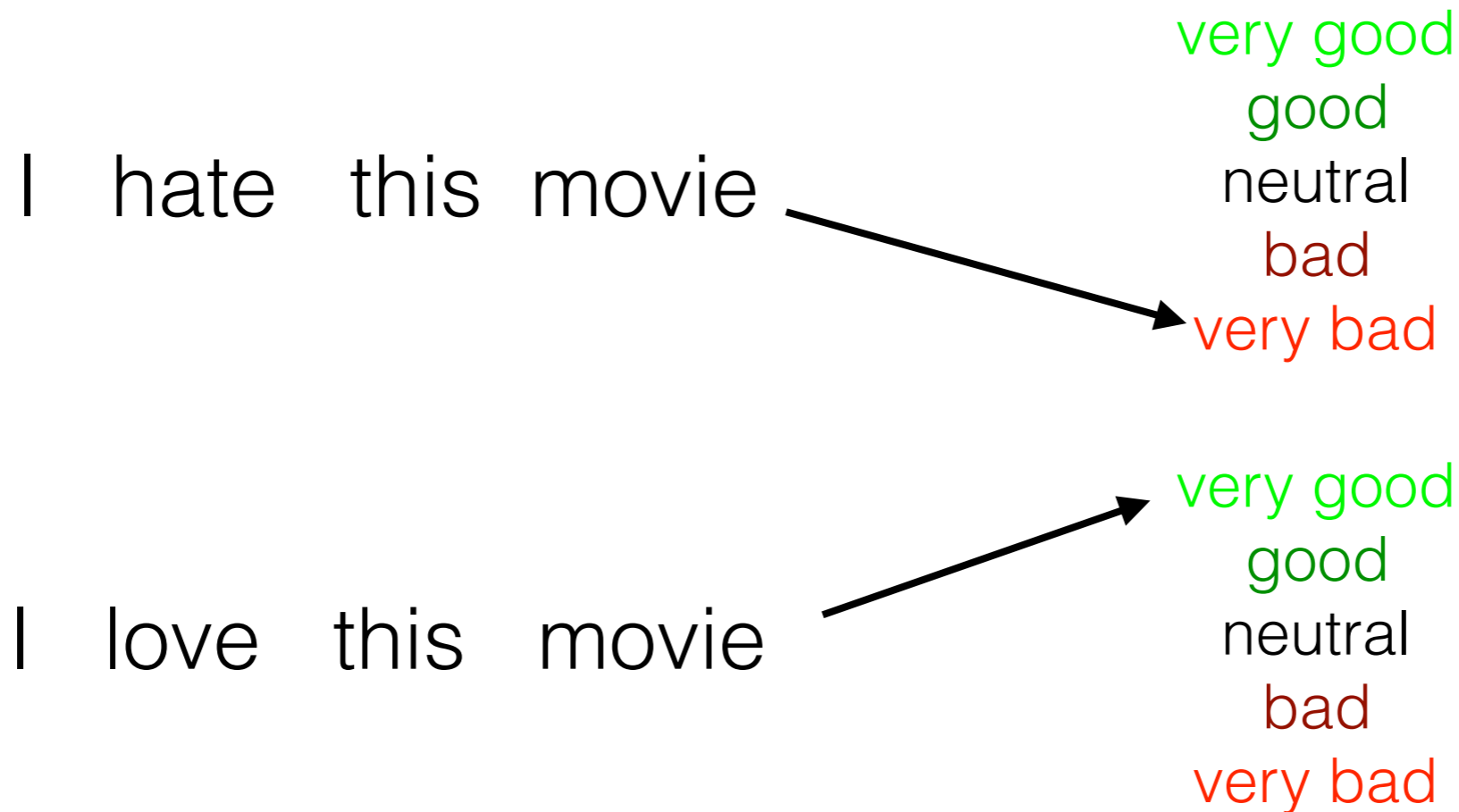
CSE 5539-0010 Ohio State University

Instructor: Wei Xu

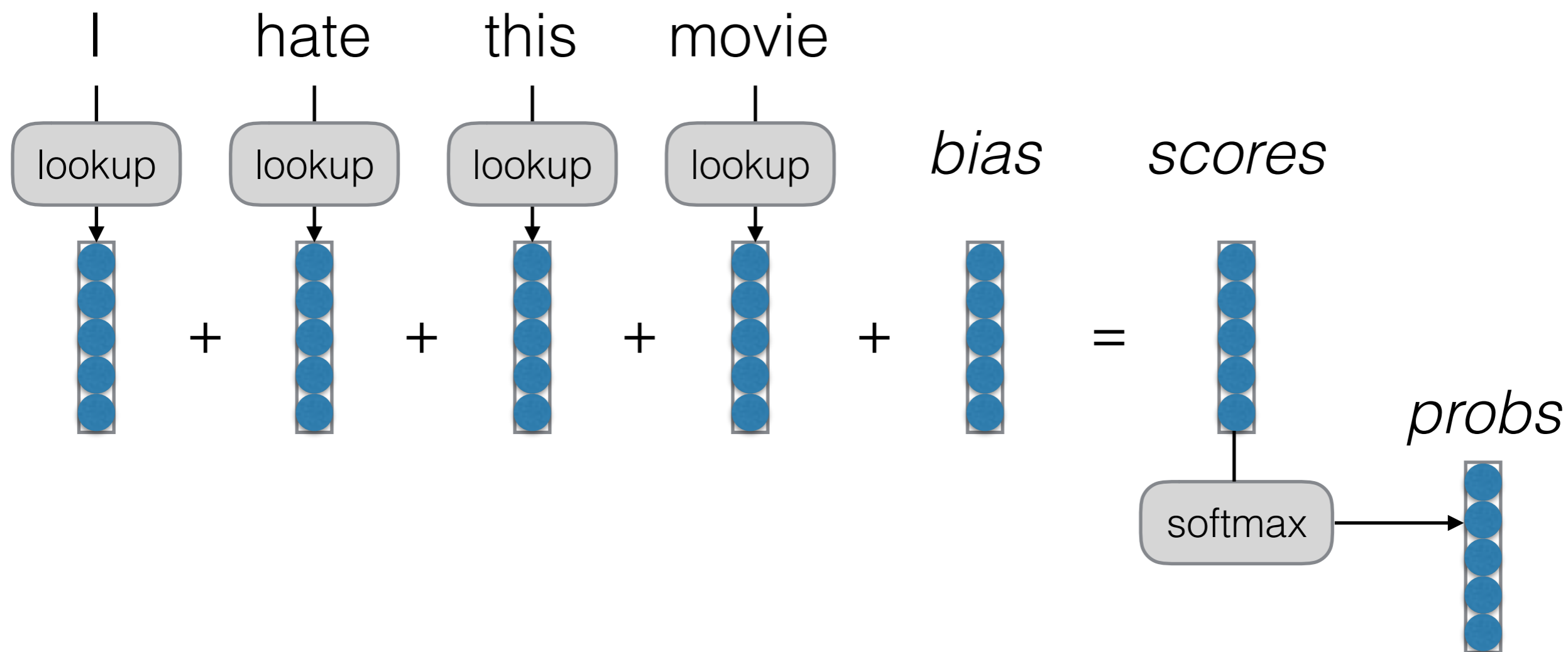
Website: socialmedia-class.org

slides are from Yang Yi, Graham Neubig, Richard Socher

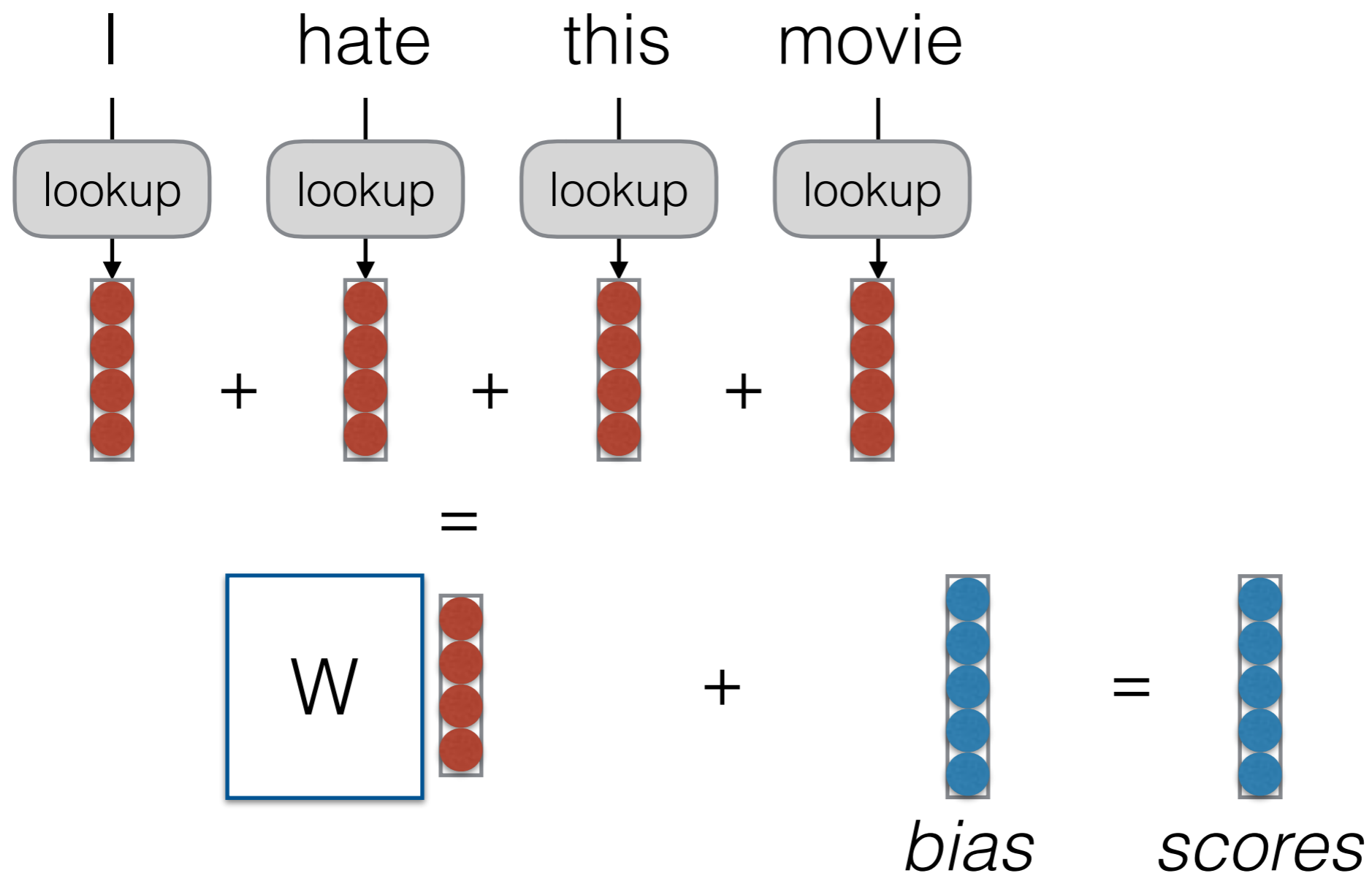
An Example Prediction Problem: Sentence Classification



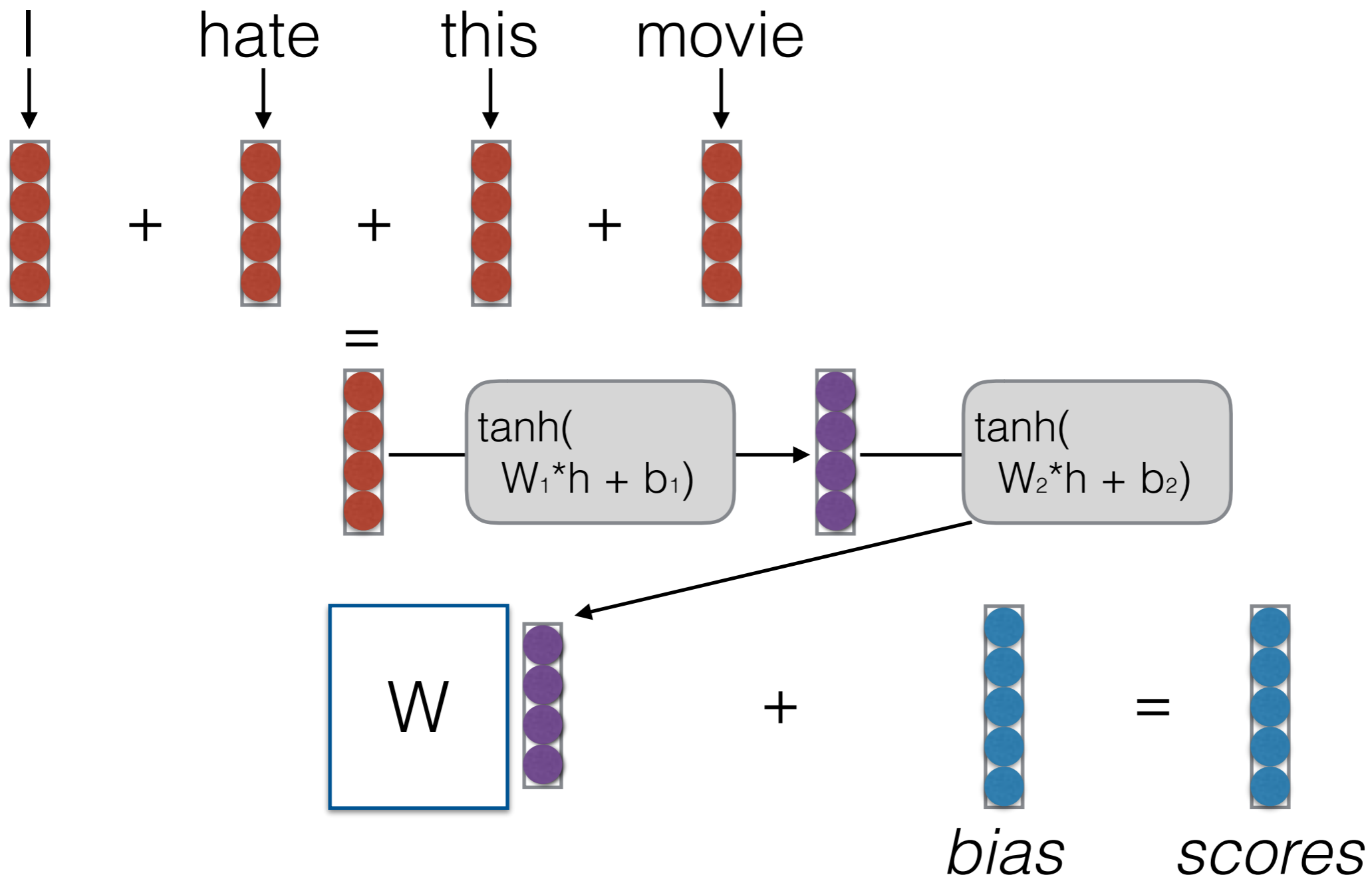
A First Try: Bag of Words (BOW)



Continuous Bag of Words (CBOW)




Deep CBOW



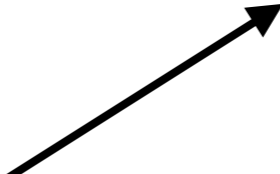
Build It, Break It

I don't love this movie



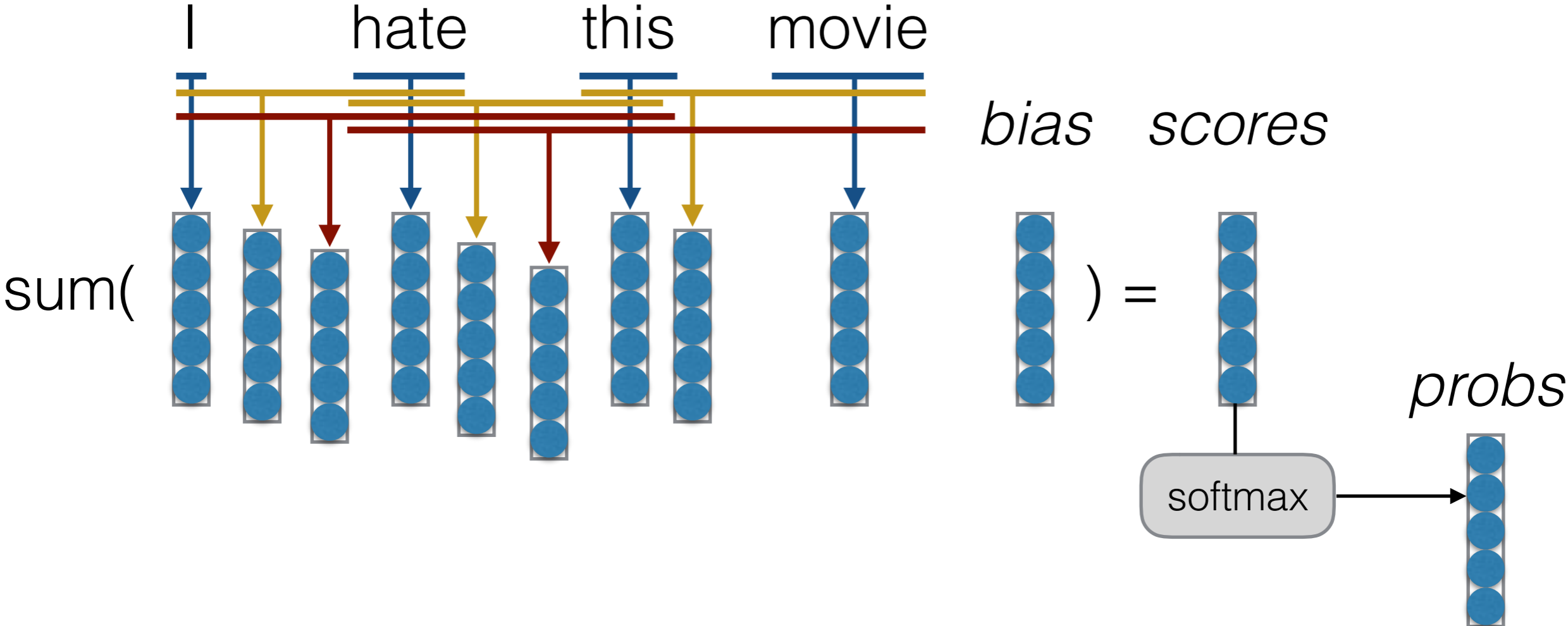
very good
good
neutral
bad
very bad

There's nothing I don't love about this movie



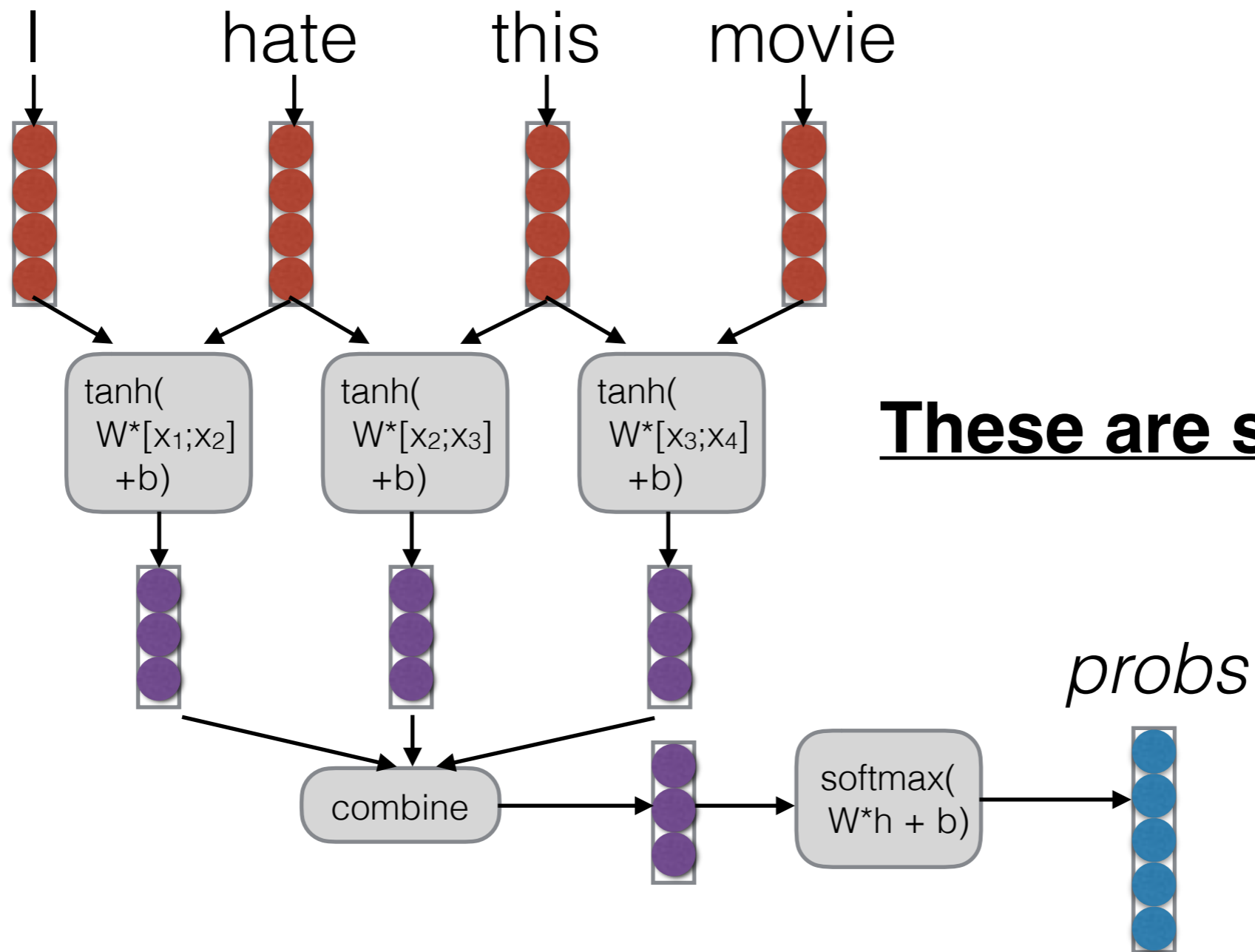
very good
good
neutral
bad
very bad

Bag of n-grams



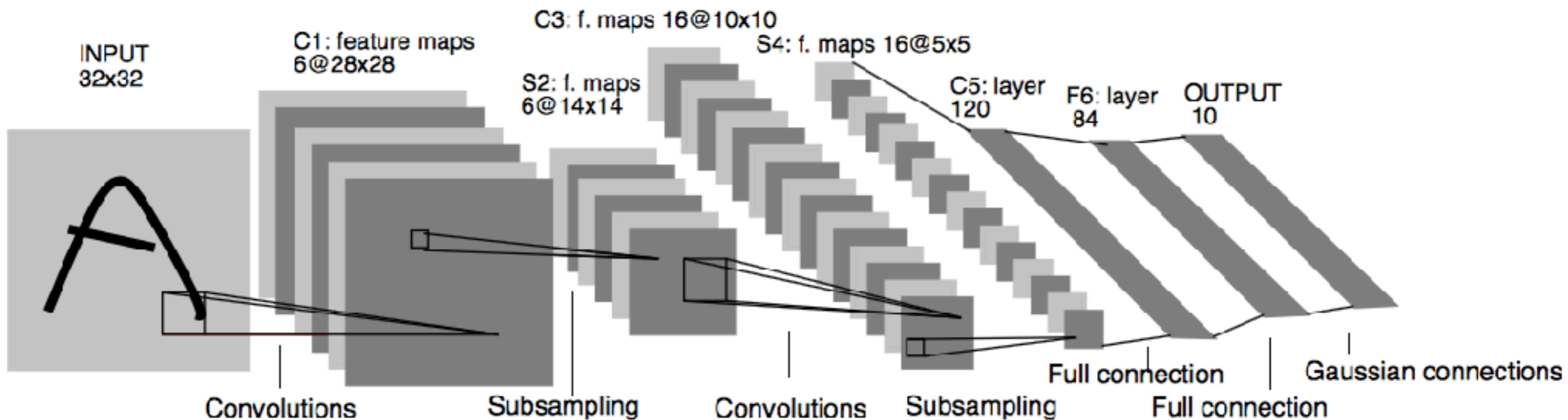
Time Delay Neural Networks

(Waibel et al. 1989)



Convolutional Networks

(LeCun et al. 1997)



Parameter extraction performs a 2D sweep, not 1D

1 _{x1}	1 _{x0}	1 _{x1}	0	0
0 _{x0}	1 _{x1}	1 _{x0}	1	0
0 _{x1}	0 _{x0}	1 _{x1}	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

Convolved
Feature

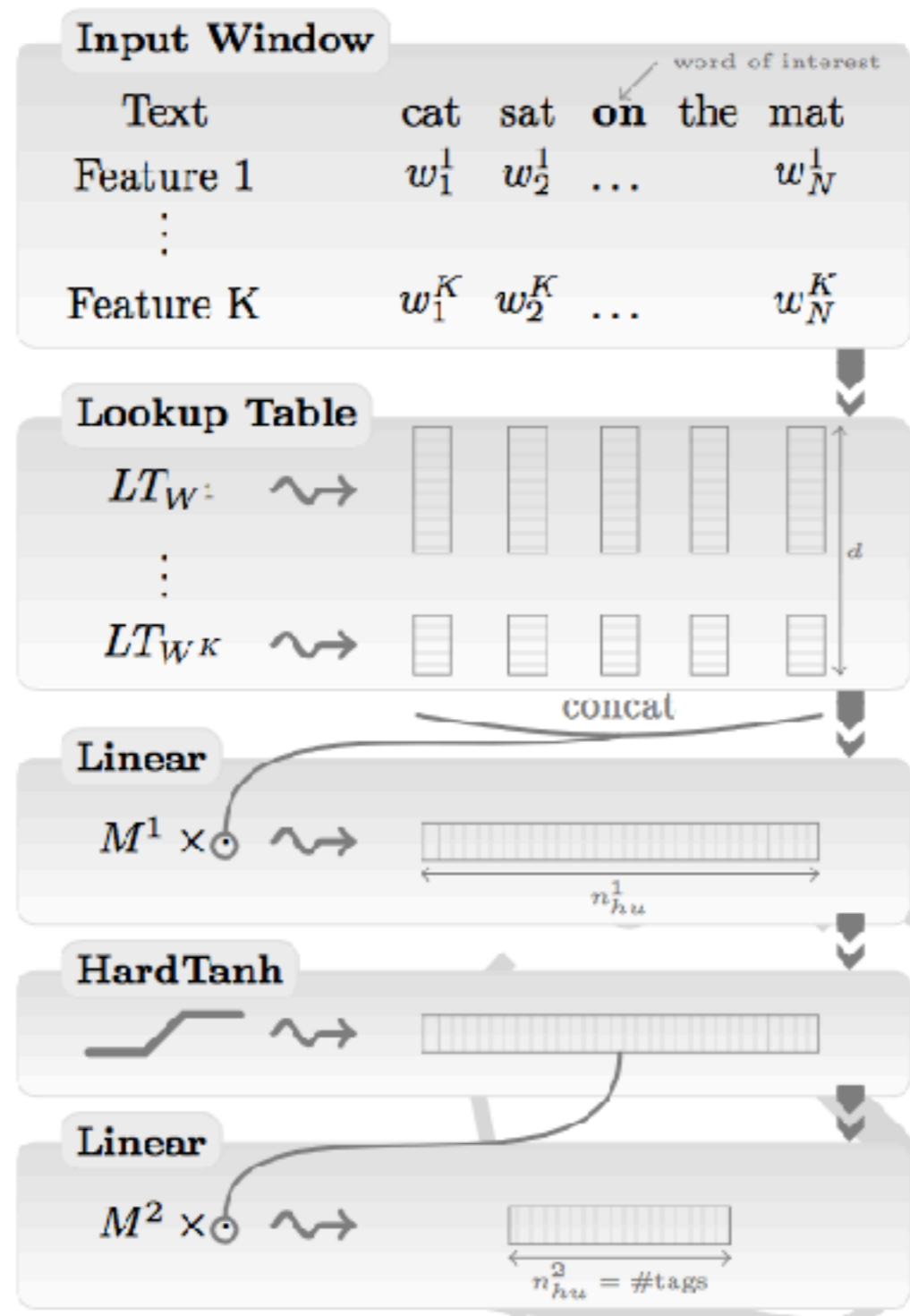
CNNs for Text

(Collobert and Weston 2011)

- 1D convolution \approx Time Delay Neural Network
 - But often uses terminology/functions borrowed from image processing
- Two main paradigms:
 - **Context window modeling:** For tagging, etc. get the surrounding context before tagging
 - **Sentence modeling:** Do convolution to extract n-grams, pooling to combine over whole sentence

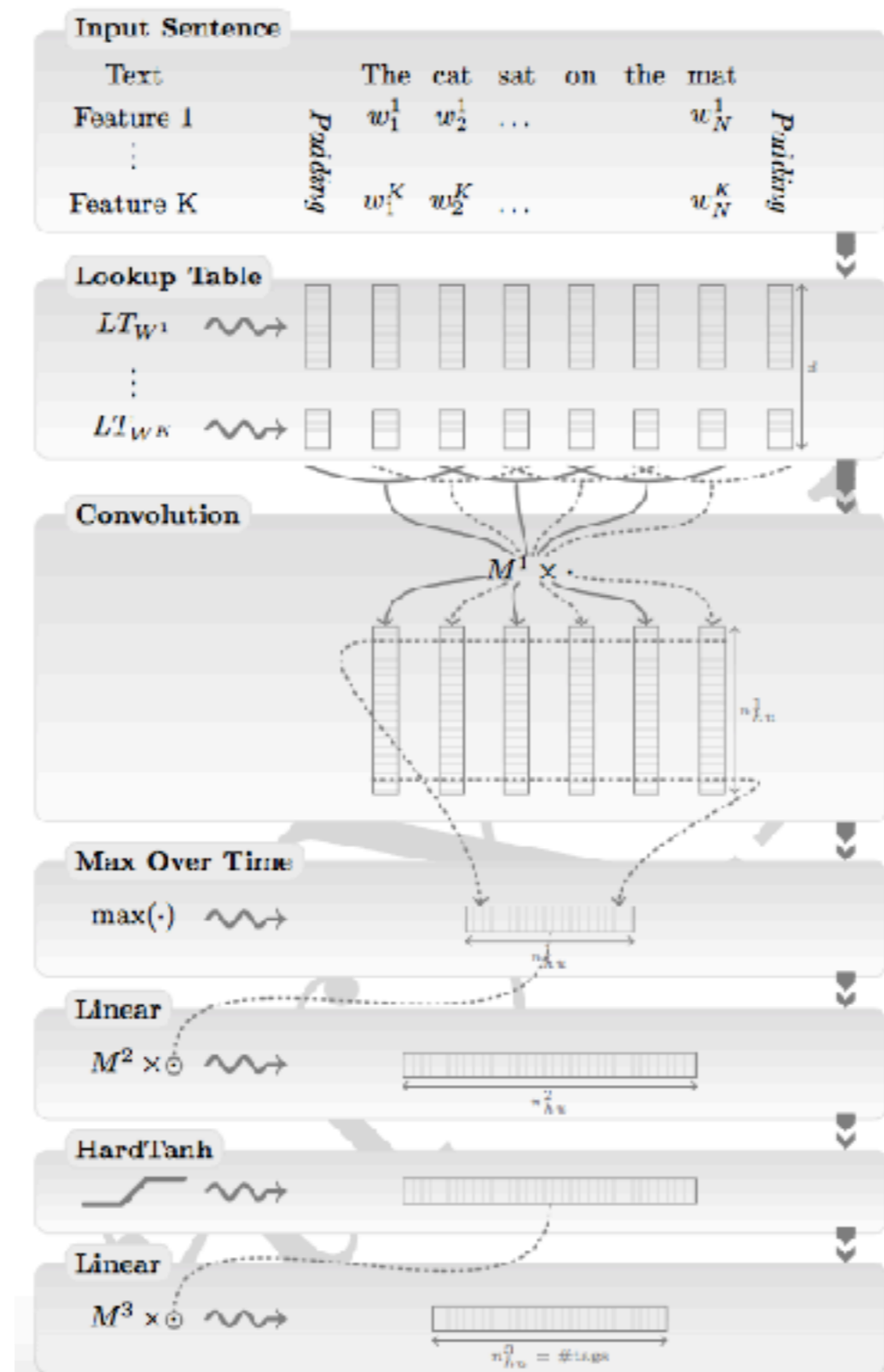
CNNs for Tagging

(Collobert and Weston 2011)



CNNs for Sentence Modeling

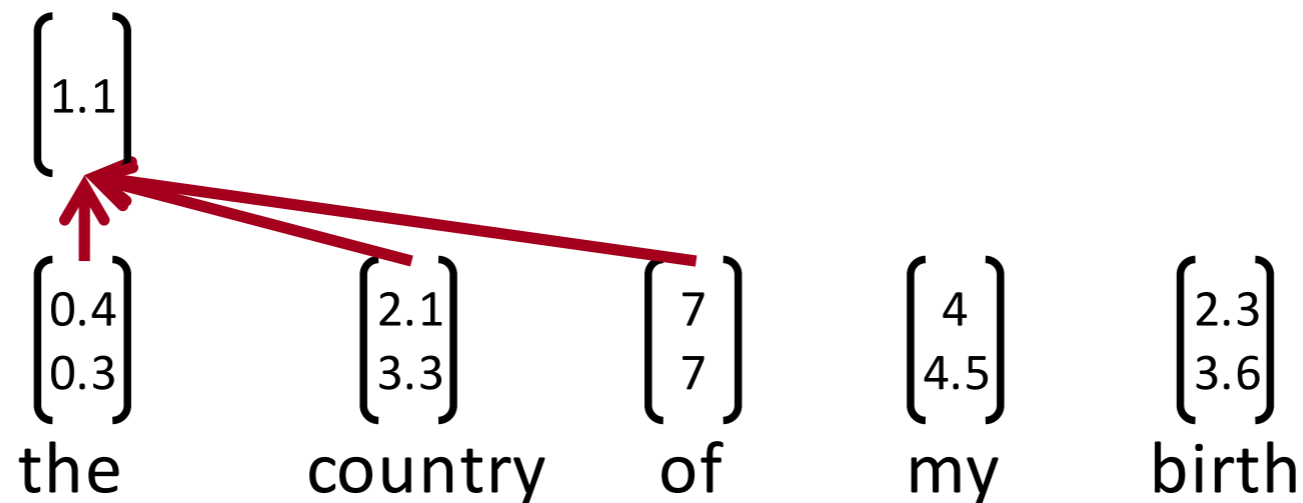
(Collobert and Weston 2011)



Single layer CNN

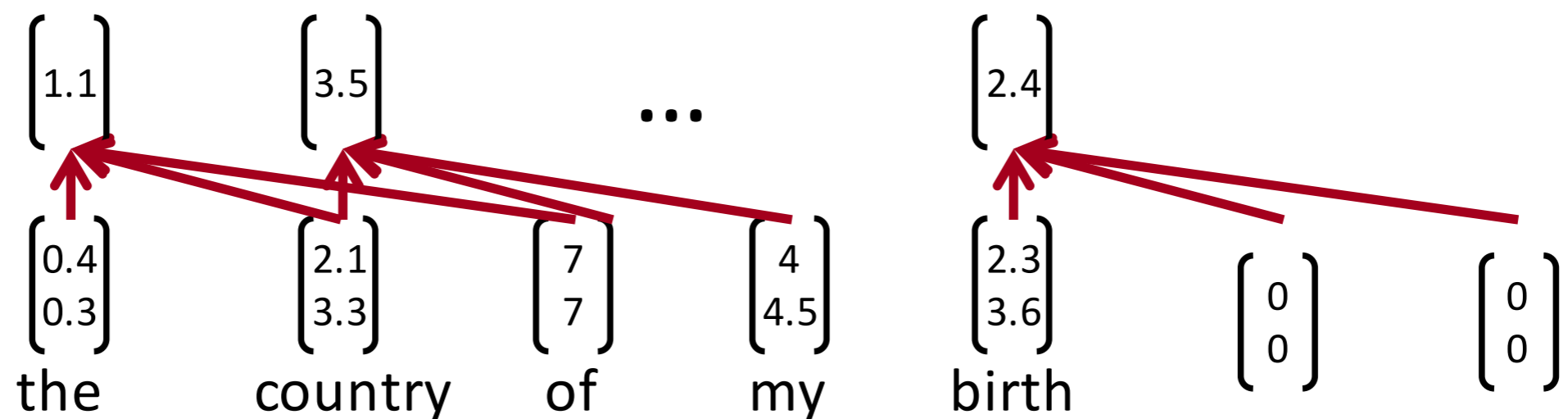
- Convolutional filter: $\mathbf{w} \in \mathbb{R}^{hk}$ (goes over window of h words)
- Note, filter is vector!
- Window size h could be 2 (as before) or higher, e.g. 3:
- To compute feature for CNN layer:

$$c_i = f(\mathbf{w}^T \mathbf{x}_{i:i+h-1} + b)$$



Single layer CNN

- Filter w is applied to all possible windows (concatenated vectors)
- Sentence: $\mathbf{x}_{1:n} = \mathbf{x}_1 \oplus \mathbf{x}_2 \oplus \dots \oplus \mathbf{x}_n$
- All possible windows of length h : $\{\mathbf{x}_{1:h}, \mathbf{x}_{2:h+1}, \dots, \mathbf{x}_{n-h+1:n}\}$
- Result is a feature map: $\mathbf{c} = [c_1, c_2, \dots, c_{n-h+1}] \in \mathbb{R}^{n-h+1}$



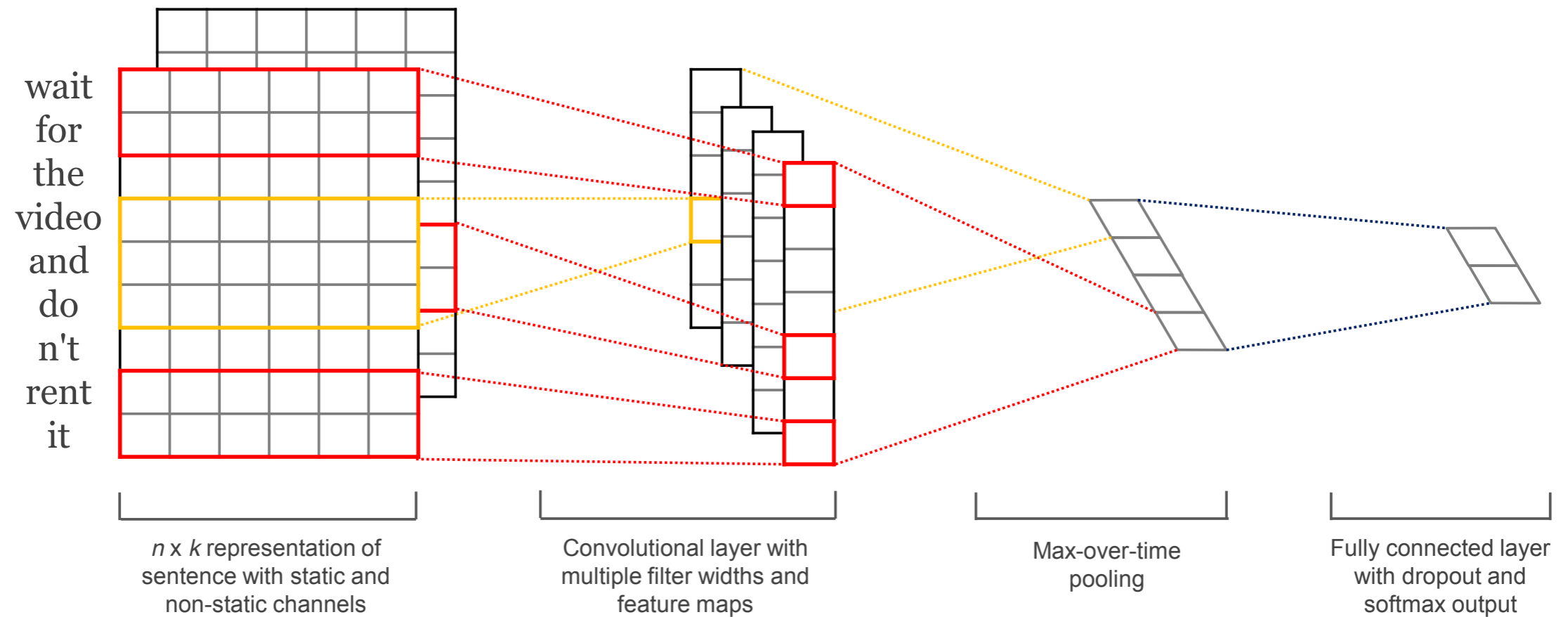
Single layer CNN: Pooling layer

- New building block: Pooling
- In particular: max-over-time pooling layer
- Idea: capture most important activation (maximum over time)
- From feature map $\mathbf{c} = [c_1, c_2, \dots, c_{n-h+1}] \in \mathbb{R}^{n-h+1}$
- Pooled single number: $\hat{c} = \max\{\mathbf{c}\}$
- But we want more features!

Solution: Multiple filters

- Use multiple filter weights w
- Useful to have different window sizes h
- Because of max pooling $\hat{c} = \max\{\mathbf{c}\}$, length of \mathbf{c} irrelevant
$$\mathbf{c} = [c_1, c_2, \dots, c_{n-h+1}] \in \mathbb{R}^{n-h+1}$$
- So we can have some filters that look at unigrams, bigrams, tri-grams, 4-grams, etc.

Figure from Kim (2014)



Tricks to make it work better: Dropout

- Idea: randomly mask/dropout/set to 0 some of the feature weights z
- Create masking vector r of Bernoulli random variables with probability p (a hyperparameter) of being 1

- Delete features during training:

$$y = \text{softmax} \left(W^{(S)} (r \circ z) + b \right)$$

- Reasoning: Prevents co-adaptation (overfitting to seeing specific feature constellations)

Tricks to make it work better: Dropout

$$y = \text{softmax} \left(W^{(S)} (r \circ z) + b \right)$$

- At training time, gradients are backpropagated only through those elements of z vector for which $r_i = 1$
- At test time, there is no dropout, so feature vectors z are larger.
- Hence, we scale final vector by Bernoulli probability p

$$\hat{W}^{(S)} = pW^{(S)}$$

- Kim (2014) reports **2 – 4% improved accuracy** and ability to use very large networks without overfitting

All hyperparameters in Kim (2014)

- Find hyperparameters based on dev set
- Nonlinearity: reLu
- Window filter sizes $h = 3, 4, 5$
- Each filter size has 100 feature maps
- Dropout $p = 0.5$
- L2 constraint s for rows of softmax $s = 3$
- Mini batch size for SGD training: 50
- Word vectors: pre-trained with word2vec, $k = 300$
- During training, keep checking performance on dev set and pick highest accuracy weights for final evaluation

A Case Study

Automatic Paraphrase Collection and Identification in Twitter

Wuwei Lan, Siyu Qiu, Hua He, Wei Xu



What is paraphrase?

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

famous
delicious
loved to eat

=

=

=

=

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

known throughout the world
tasty
enjoyed eating

Paraphrase Application



Search

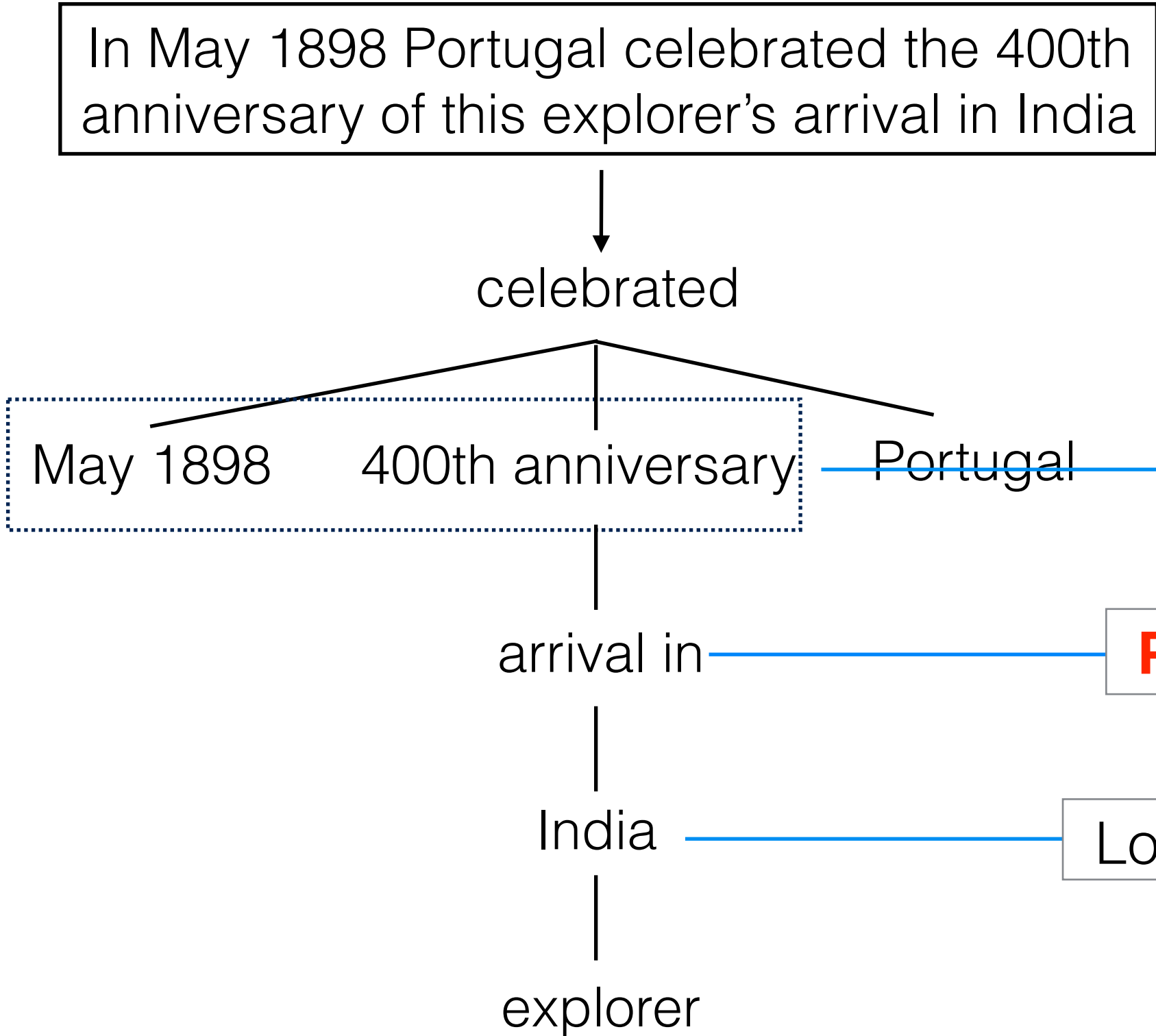
search

2477 votes	38 answers	<div><h3>Q: Sort a Python dictionary by value</h3><p>, 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 ...</p><div><div>python</div><div>sorting</div><div>dictionary</div></div><div>asked Mar 5 '09 by Gern Blanston</div></div>
12 votes	2 answers	<div><h3>Q: How to sort a Python dictionary by value?</h3><div>[duplicate]</div></div>
-4 votes	3 answers	<div><h3>Q: Python how to sort a dictionary by value in reverse order</h3><div>[duplicate]</div></div>

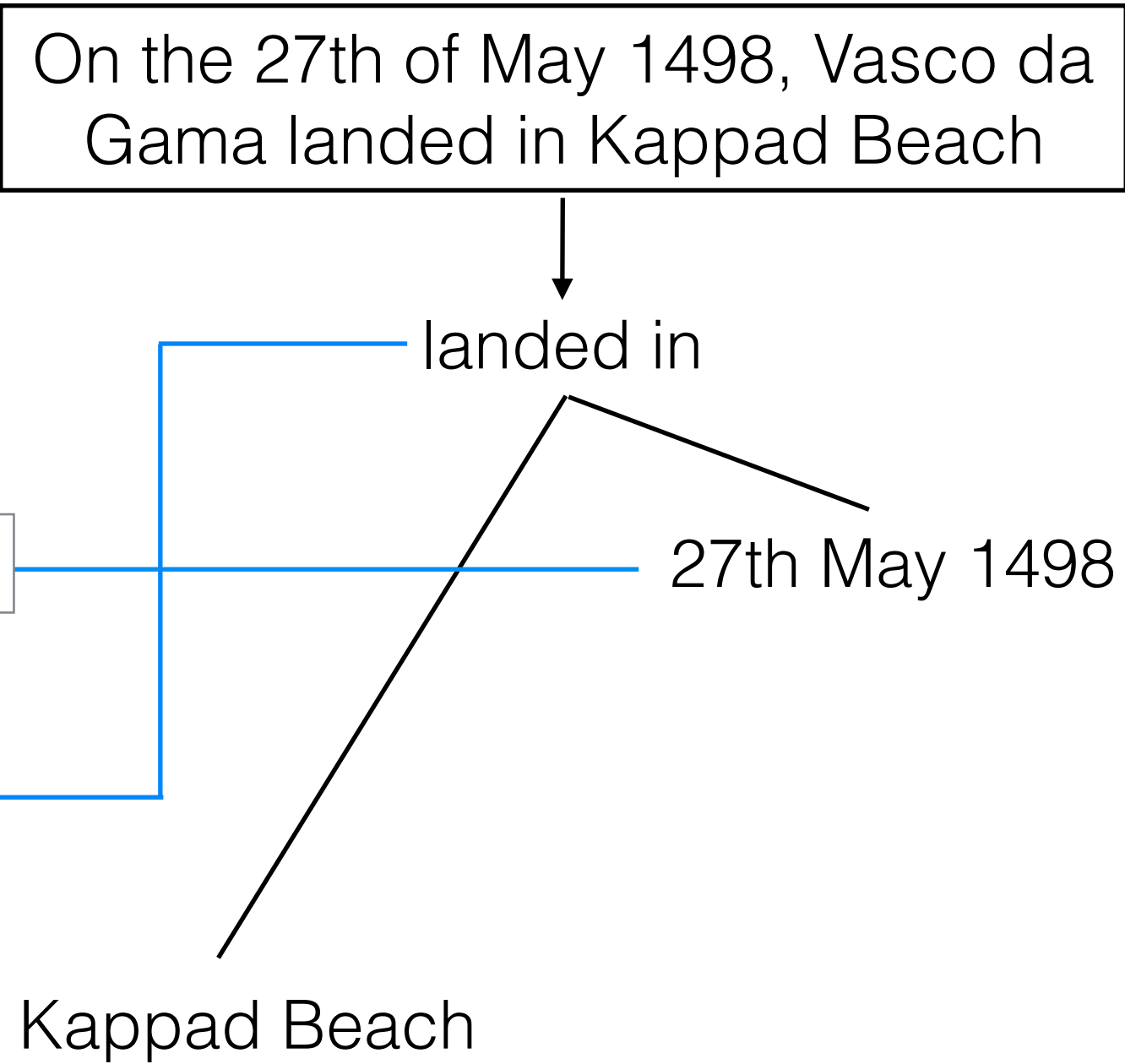
Paraphrase Application



[Question]



[Supporting Evidence]



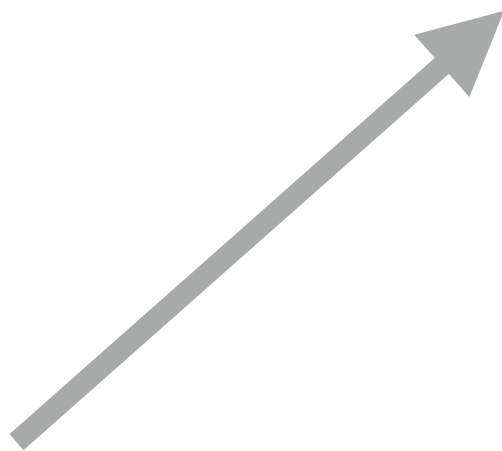
Date Match



Paraphrase

Location Match

Paraphrases?

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

[illegible]

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

Paraphrases?

<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







Paraphrase

Paraphrases?

<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


 0

 0

 0

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

 0

 0


 0

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

5 261 144



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

1 0 0



Herbert Buchsbaum  @herbertnyt · 12 Oct 2016
New bulletin from Thai palace: King is still on a ventilator and in unstable condition. [nyti.ms/2dW1A37](https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html)


1 0 0



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](https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html)
5 261 144

 **Career Synchronicity** @careersync_now · 12 Oct 2016
Fears for King's Health Shake Thailand
[ift.tt/2d7frGd](https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html)

 **Herbert Buchsbaum**  @herbertnyt · 12 Oct 2016
New bulletin from Thai palace: King is still on a ventilator and in unstable condition.
[nyti.ms/2dW1A37](https://www.nytimes.com/2016/10/13/world/asia/thailand-king.html)

same URL

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**

[1] Dolan et al., 2004

[2] Xu et al., 2014

[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)

→ ↺

Twitter, Inc. [US]https://twitter.com/search?q=Trailer&src=tren

🔍 ☆ 📷 📺 📄 📱

🏠 Home ⚡ Moments 🔔 Notifications ✉ Messages 🐦 Trailer 🔍 👤 📌

Germany Trends · [Change](#)

[#1WortRuiniertDenFilm](#)

[#DuSchlingel](#)

[#Frankfurtfilme](#)

[#bananaberlin](#)

[#Niklas](#)

[Wort Europa](#)

[Trailer](#)


[Bargeld](#)


[Nachwuchs](#)


[Maizière die Hand](#)

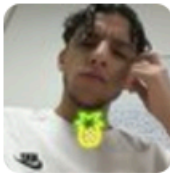
© 2017 Twitter About Help Center Terms Privacy policy Cookies Ads info



10 new results

**Gunshow Gov** @HomoHulk · 2m
Replying to @Aftashok
There's a **trailer**?
↩ 1 ↺ ❤

**Jason Blundell** @_JasonBlundell · 2m
The song for dlc5's **trailer** should be the boys are back in town
↩ ↺ 1 ❤ 1

**Kei Casi** @linuen · 3m
I can't handle [#Defenders](#)!!! So much awesomesauce in one **trailer**! I kent!
↩ ↺ ❤

**zoro** @achkamui · 3m
The DEFENDERS **Trailer** 🤯🤯🤯🤯
↩ ↺ ❤

**Pink Spoons** @pink_spoons · 3m
Check out the Dark Tower **trailer** here: bit.ly/2qrGt0P
And here's the **trailer** for The Defenders: bit.ly/2pHr3og


[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: 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**

[1] Dolan et al., 2004

[2] Xu et al., 2014

[PIT-2015_[2]]

Twitter
trending topics

14,035 annotated pairs

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

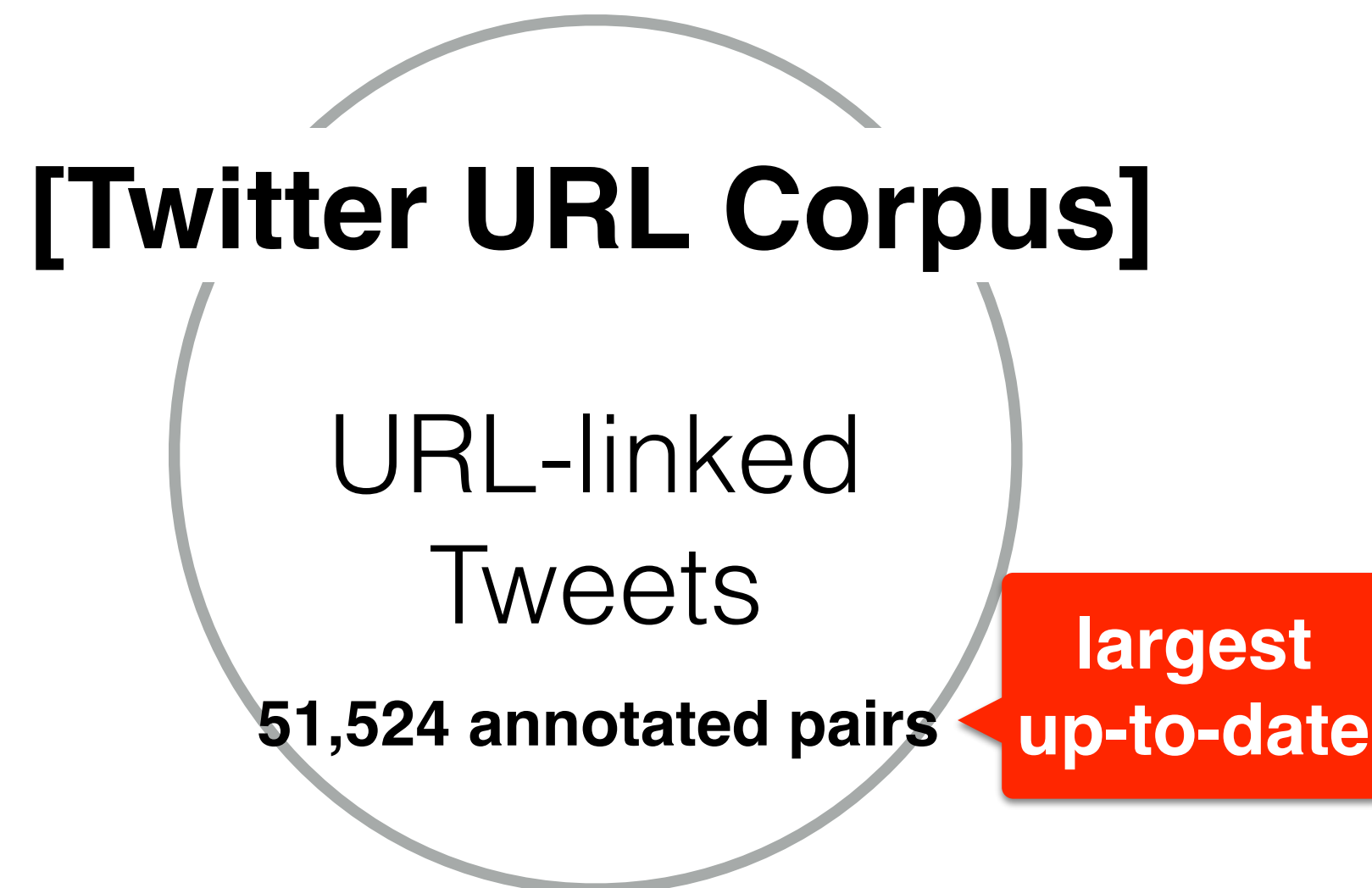
We created the 3rd paraphrase corpora (largest annotated corpus to date)

Key for success:

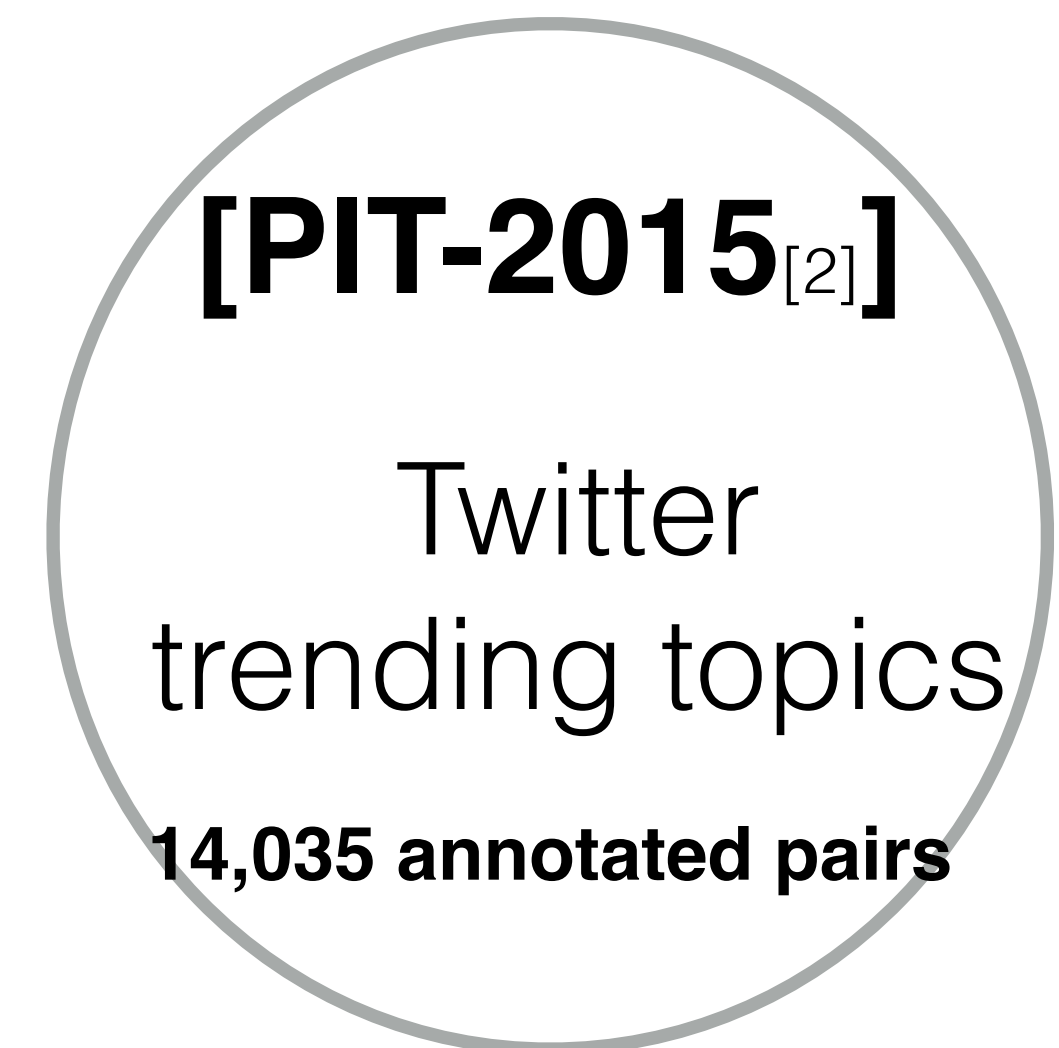
- narrow the search space
- ensure diversity among sentences
- **the simpler the better!**



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



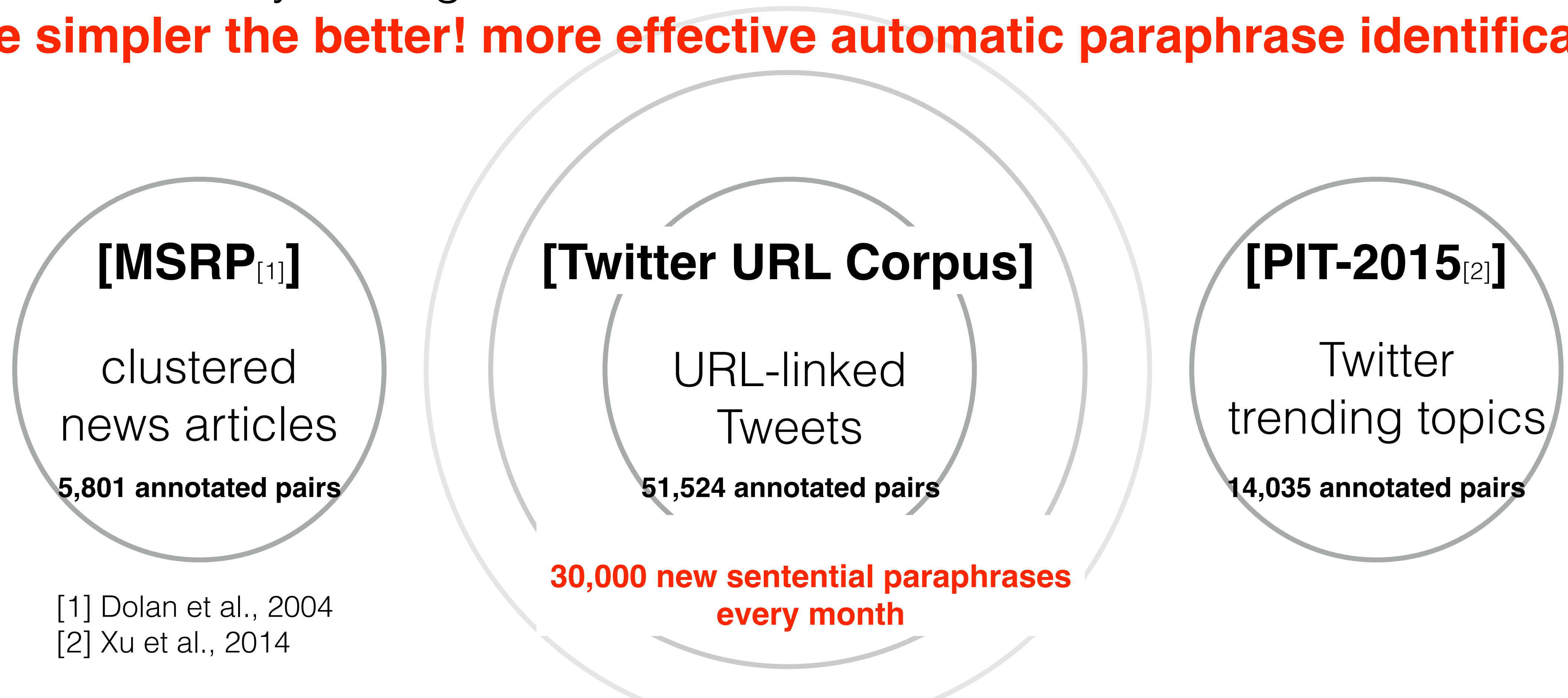
no clustering or topic detection needed
no data selection steps needed



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**



**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 @realDonaldTrump, 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

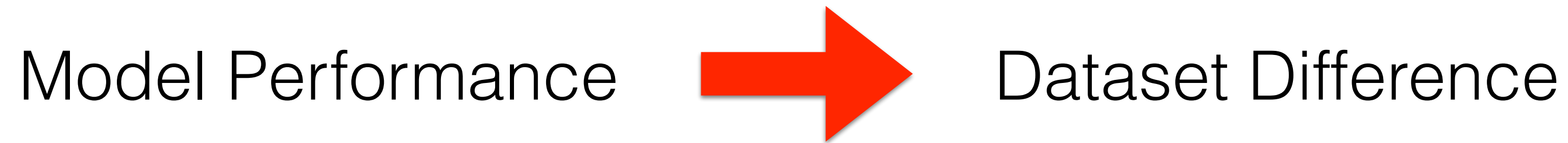
FBI supports CIA assertion

FBI Clapper back CIA's view

The FBI backs the CIA's assessment

FBI Backs CIA ...

How different from existing paraphrase corpora?

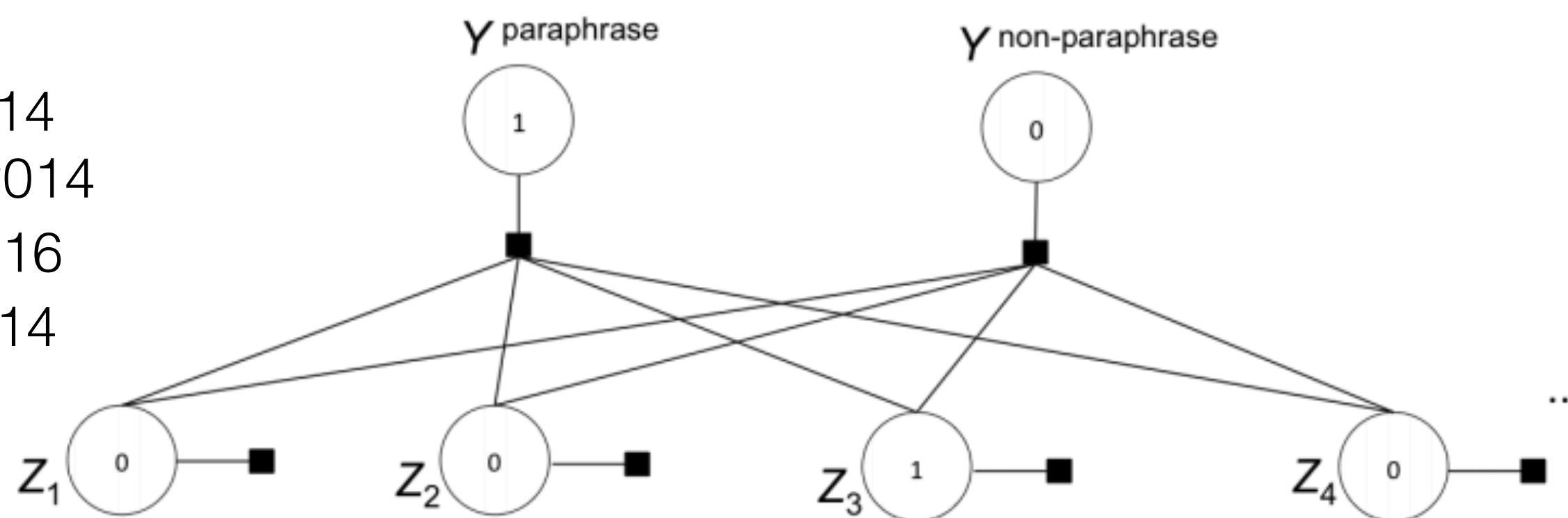


Automatic Paraphrase Identification

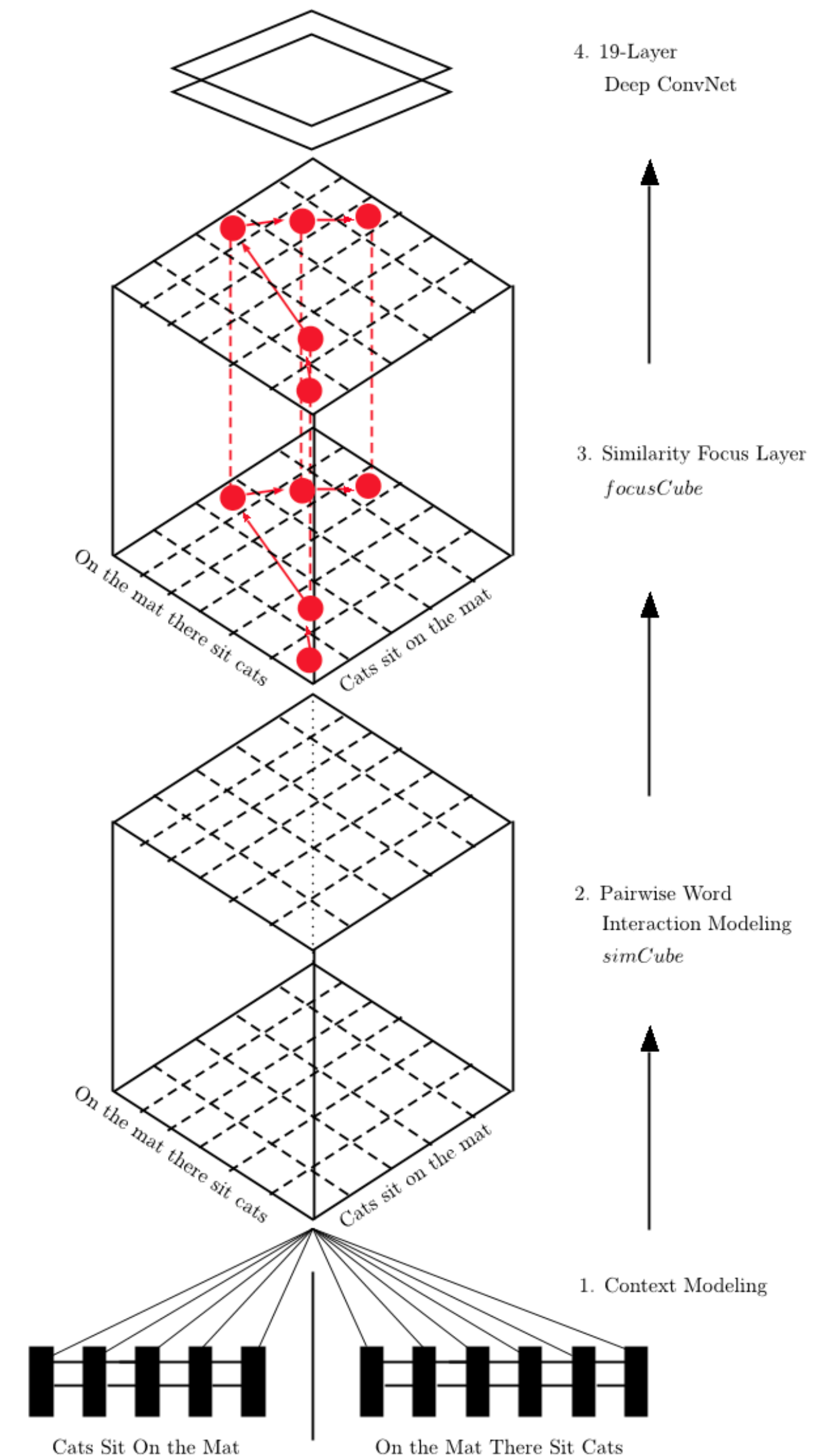
$$X \approx P \times Q^T$$

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

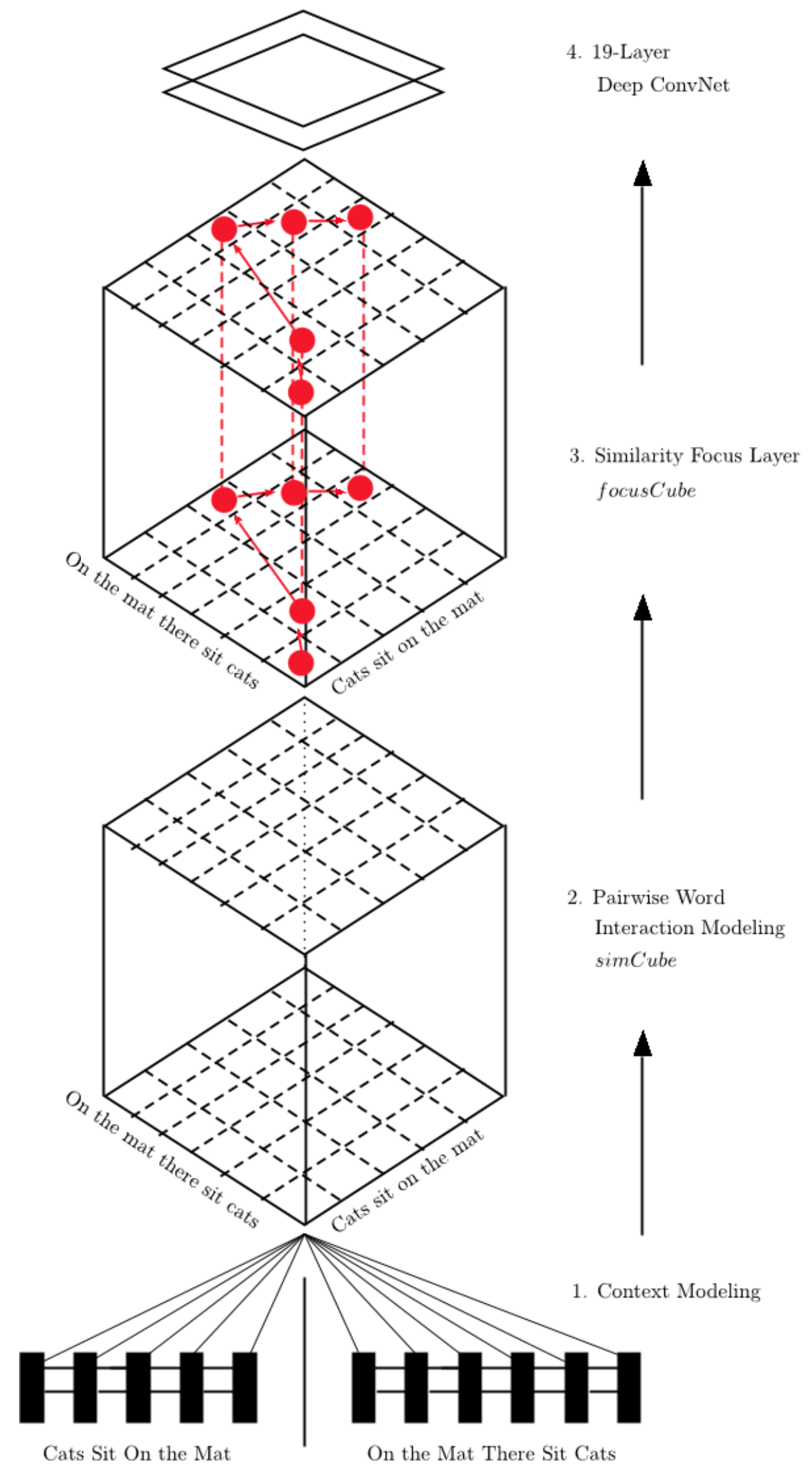
- [1] Xu et al., 2014
 [2] Guo et al., 2014
 [3] He et al., 2016
 [4] Xu et al., 2014



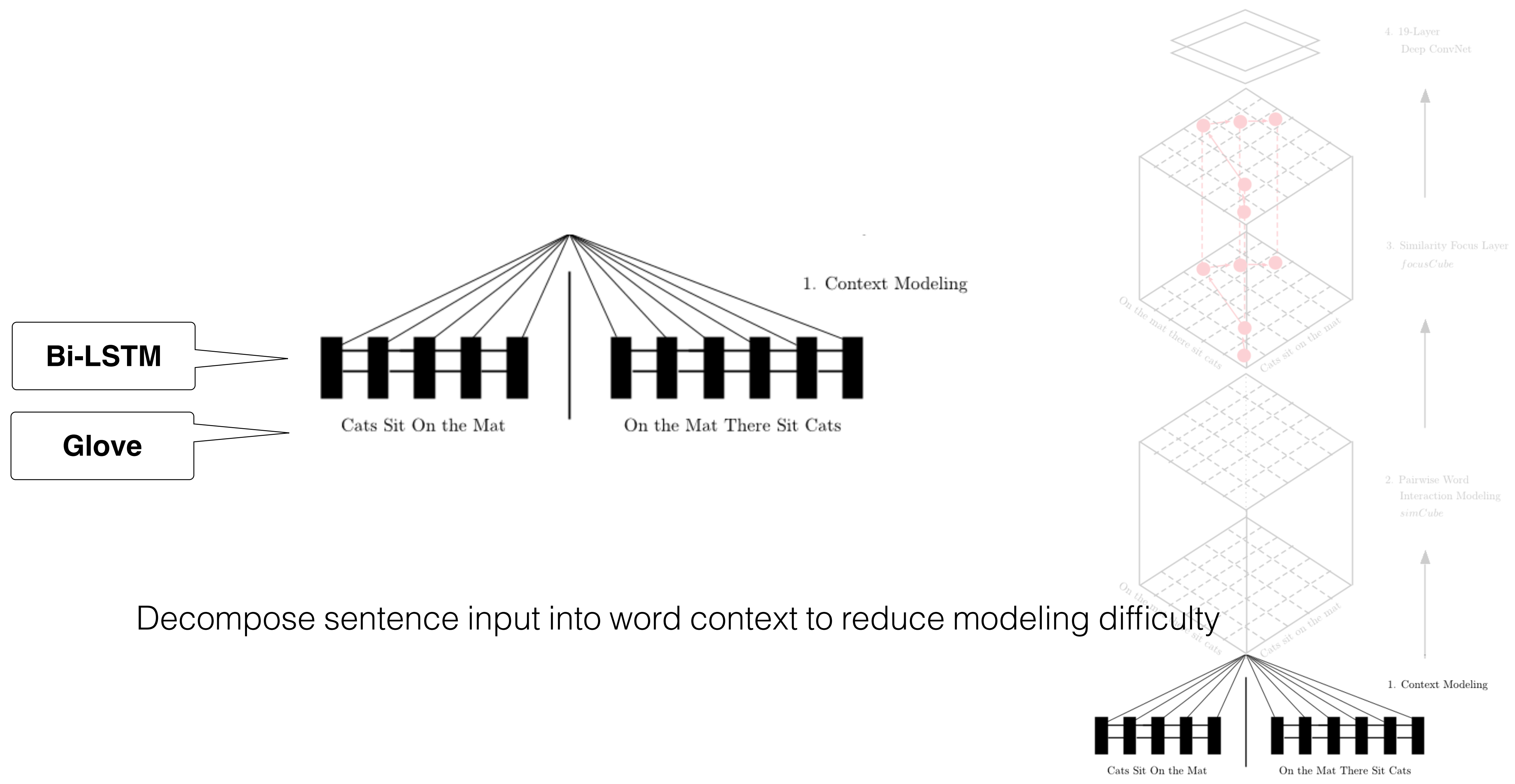
$$P(\mathbf{z}_i, y_i | \mathbf{w}_i; \theta) = \prod_{j=1}^m \exp(\theta \cdot f(z_j, w_j)) \times \sigma(\mathbf{z}_i, y_i)$$



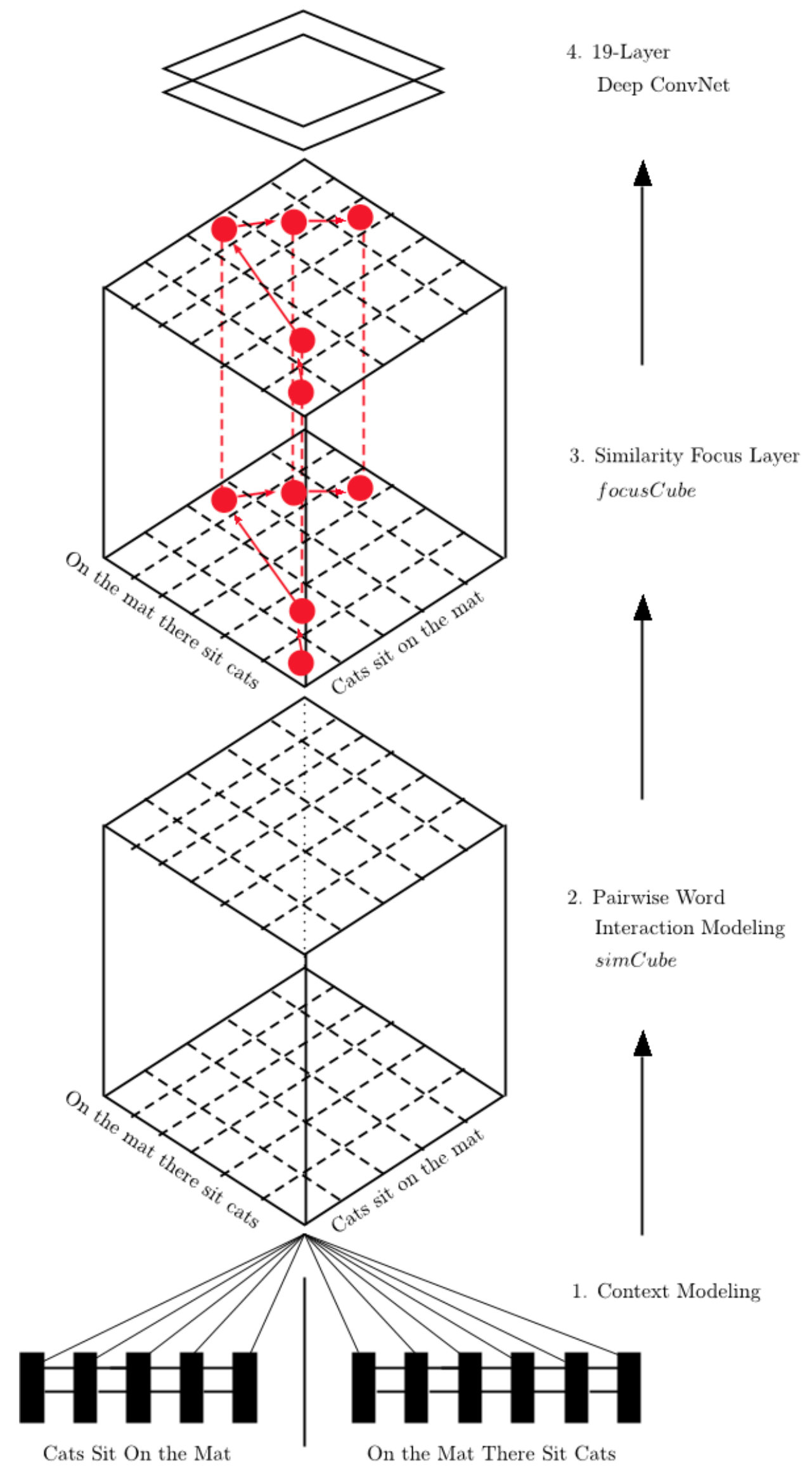
Deep Pairwise Word Model



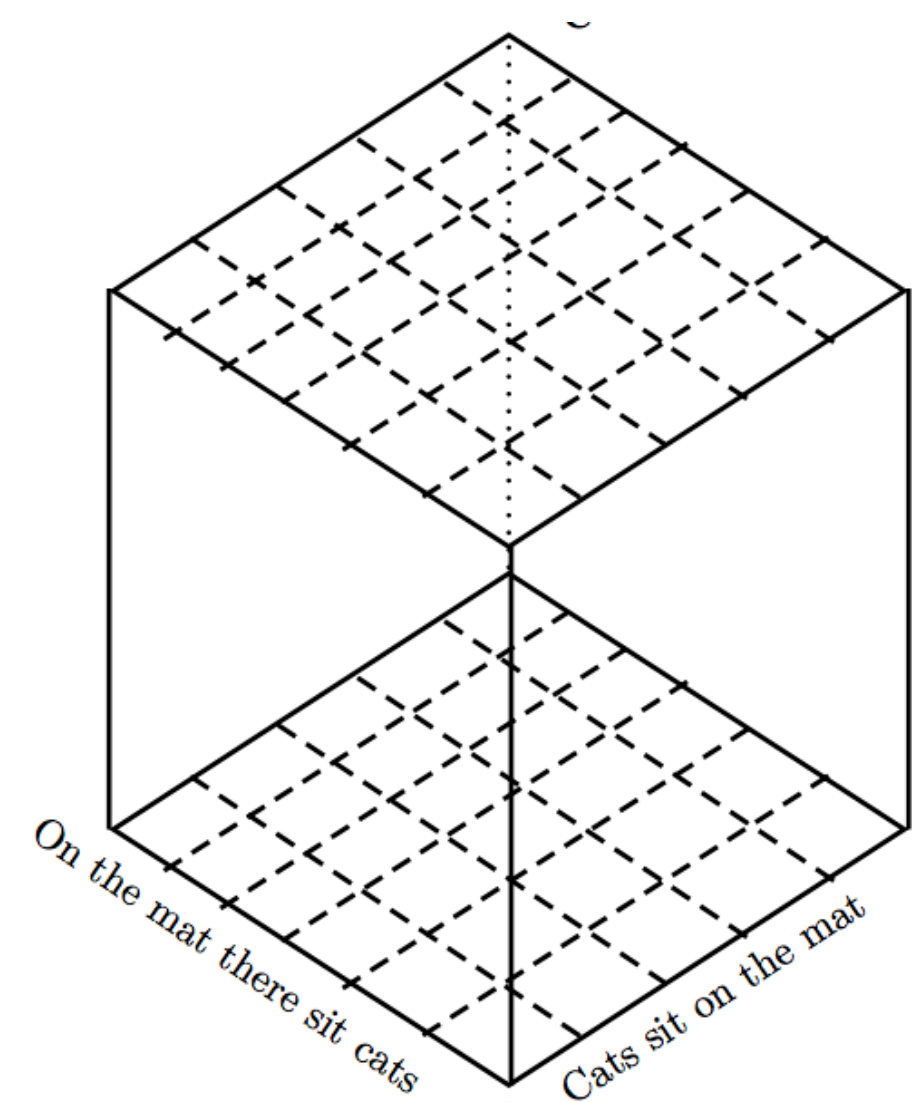
Deep Pairwise Word Model



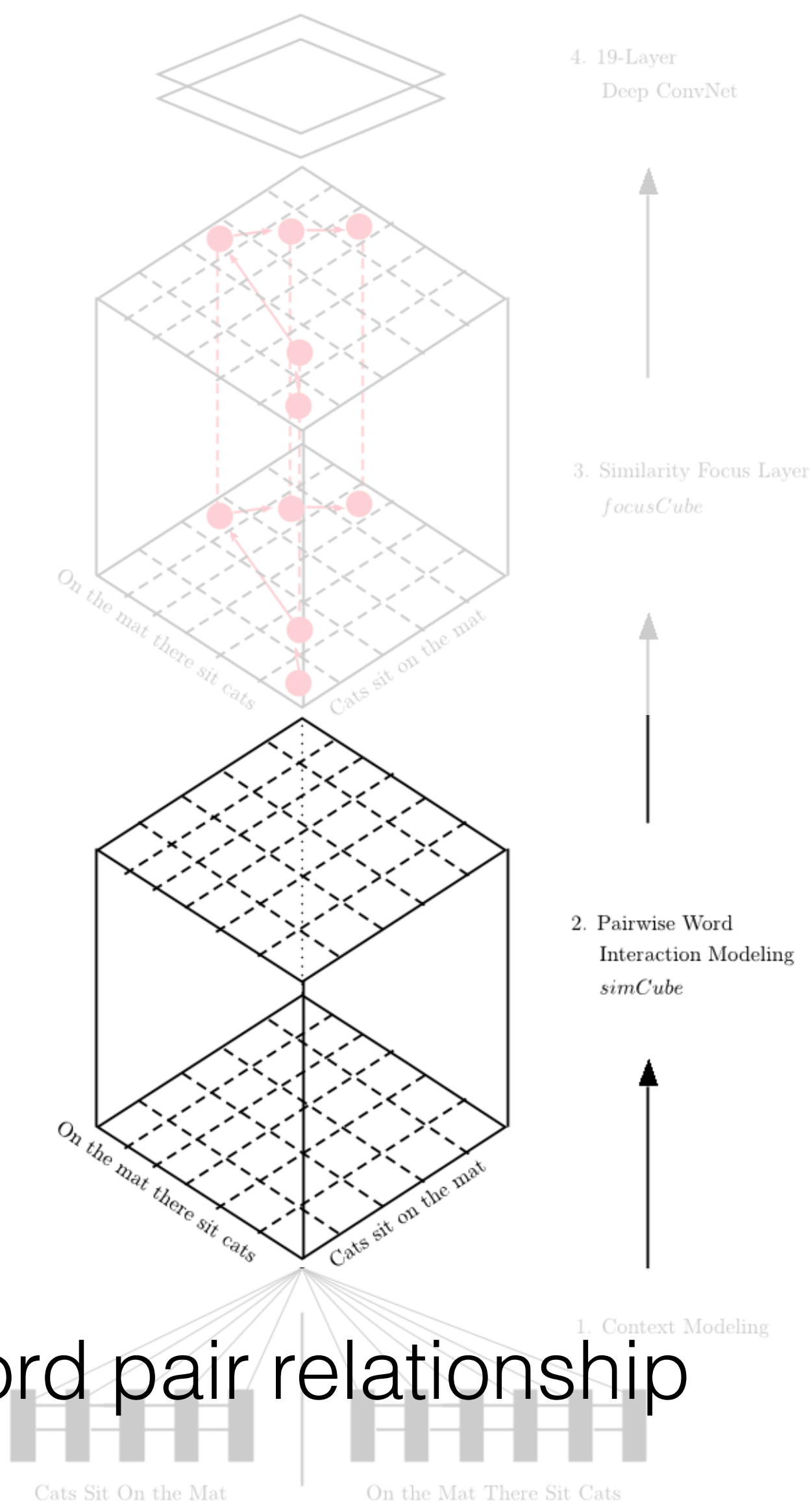
Deep Pairwise Word Model



Deep Pairwise Word Model

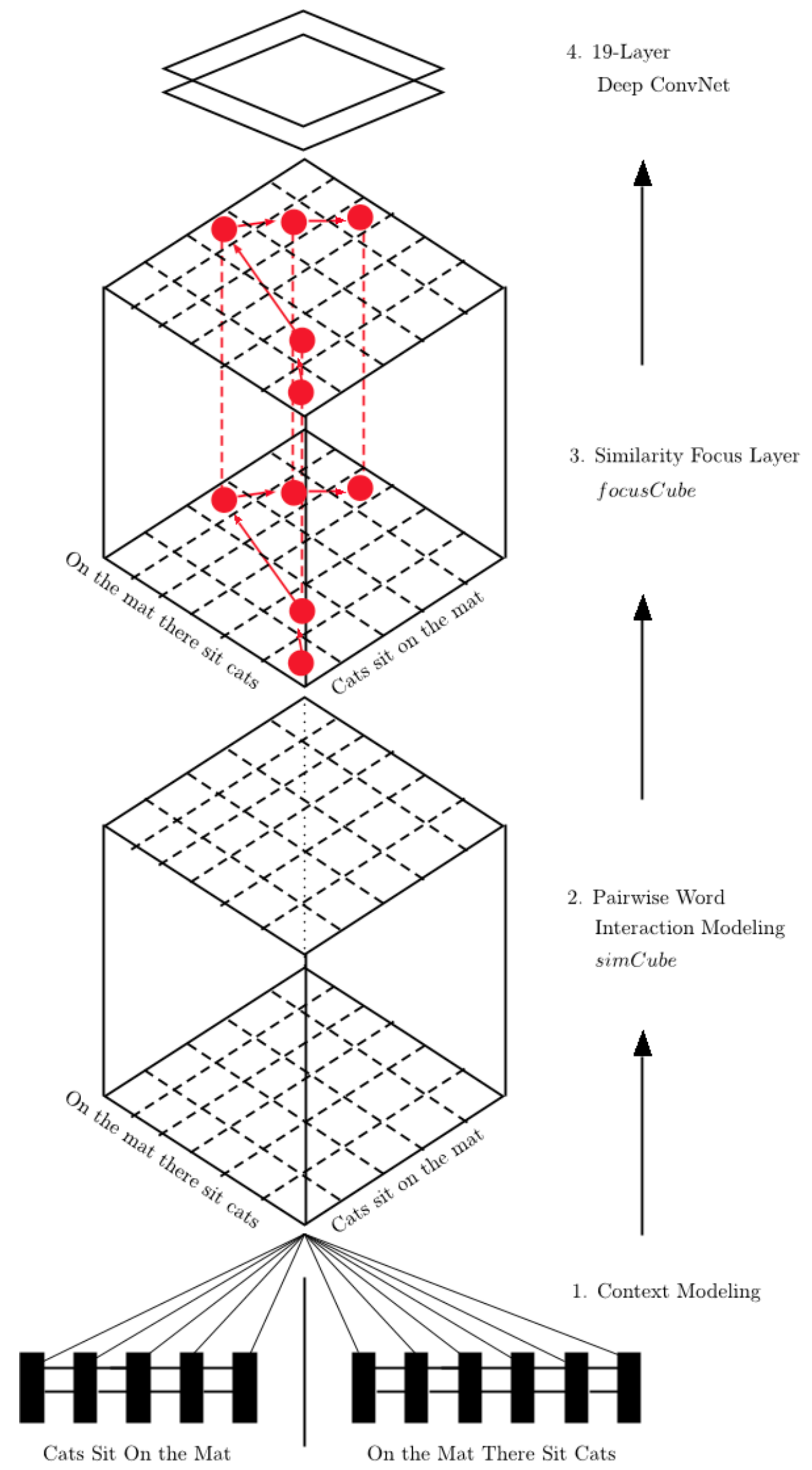


$$coU(\vec{h}_1, \vec{h}_2) = \{\cos(\vec{h}_1, \vec{h}_2), L_2 Euclid(\vec{h}_1, \vec{h}_2), \\ DotProduct(\vec{h}_1, \vec{h}_2)\}$$

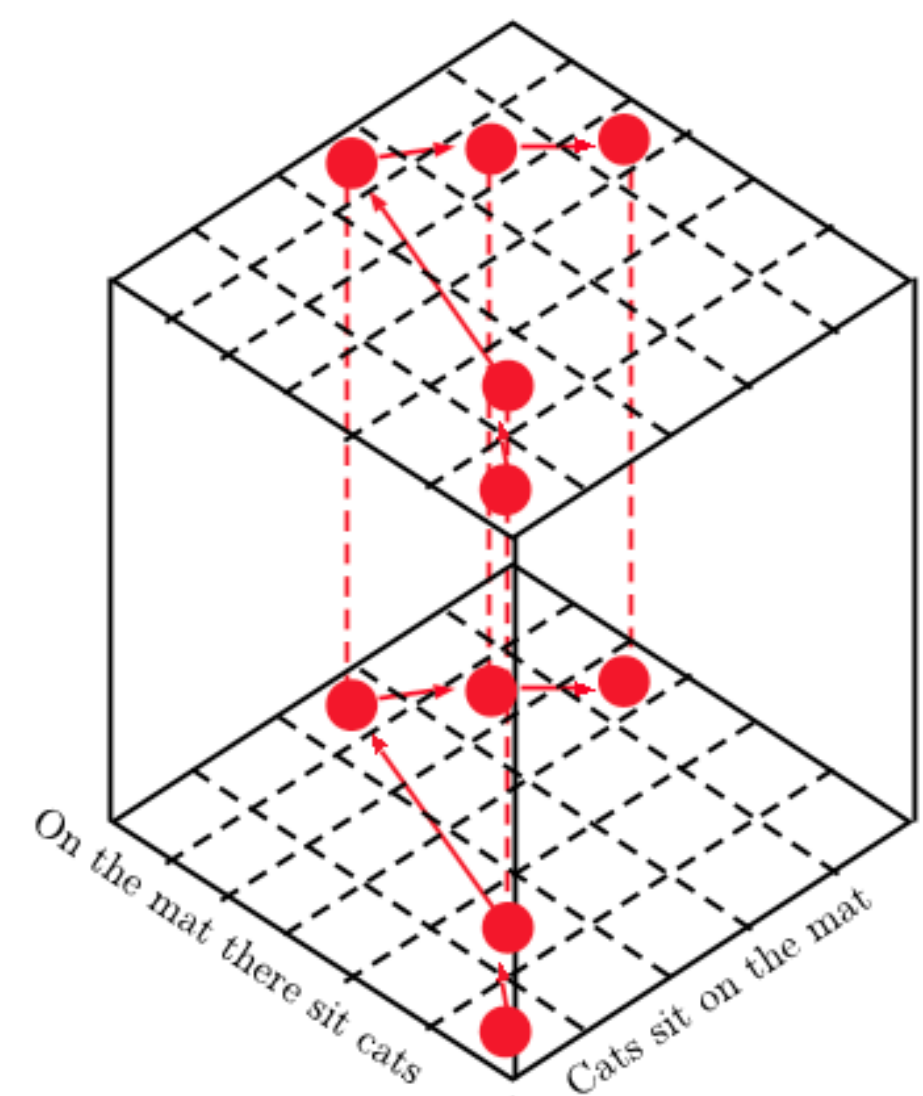


Multiple vector similarity measurement used to capture word pair relationship

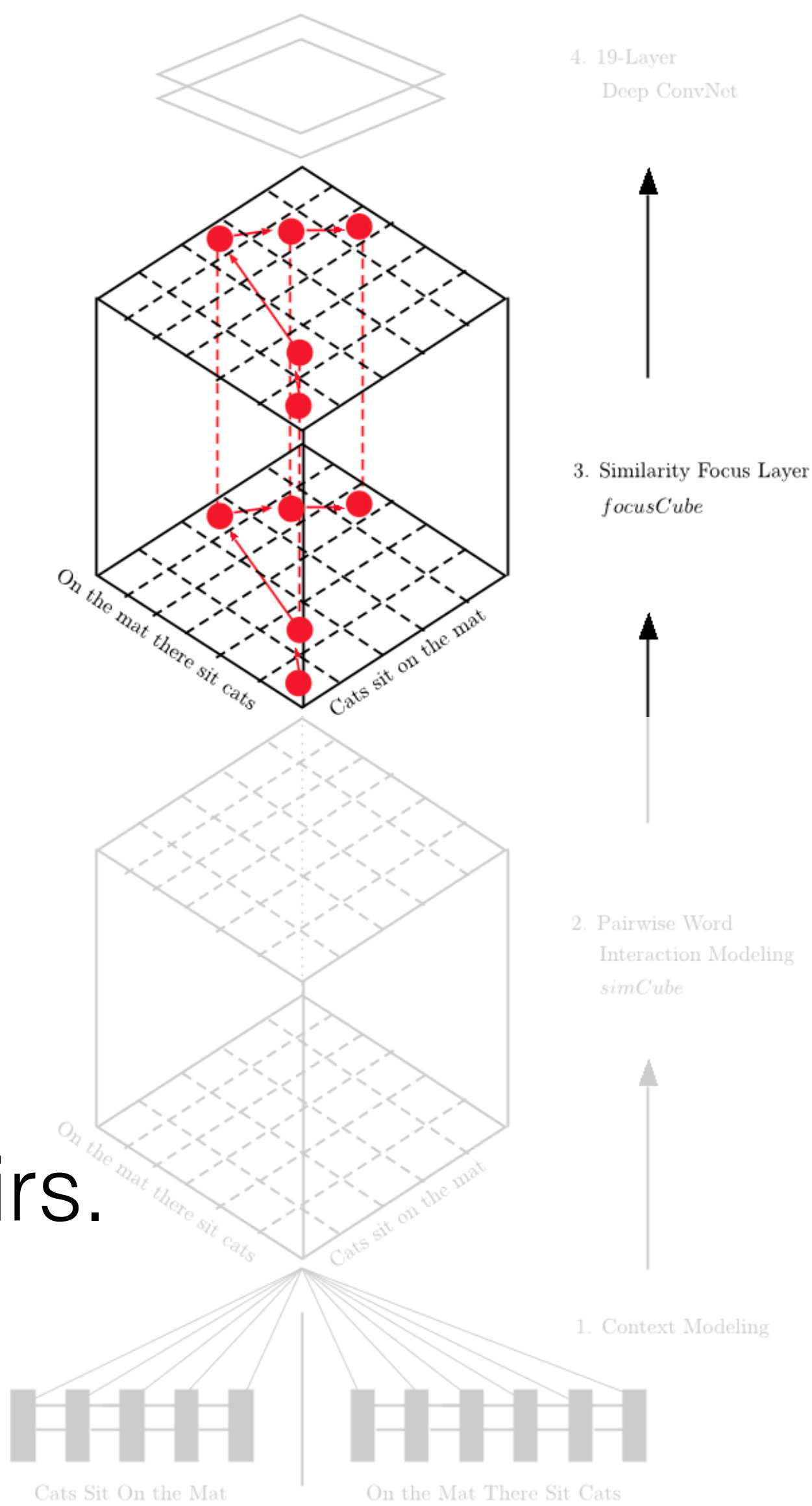
Deep Pairwise Word Model



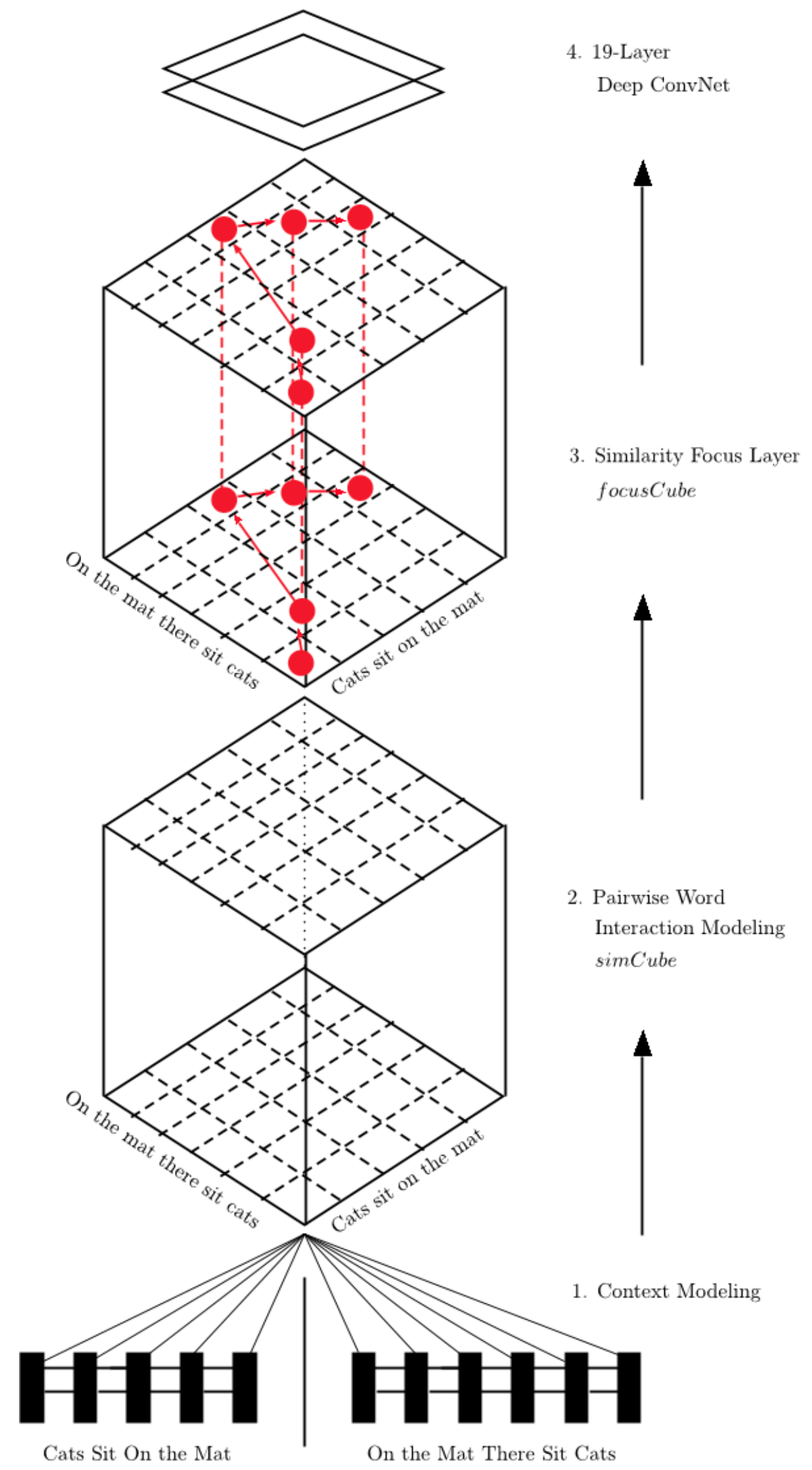
Deep Pairwise Word Model



More attention added to top ranked word pairs.



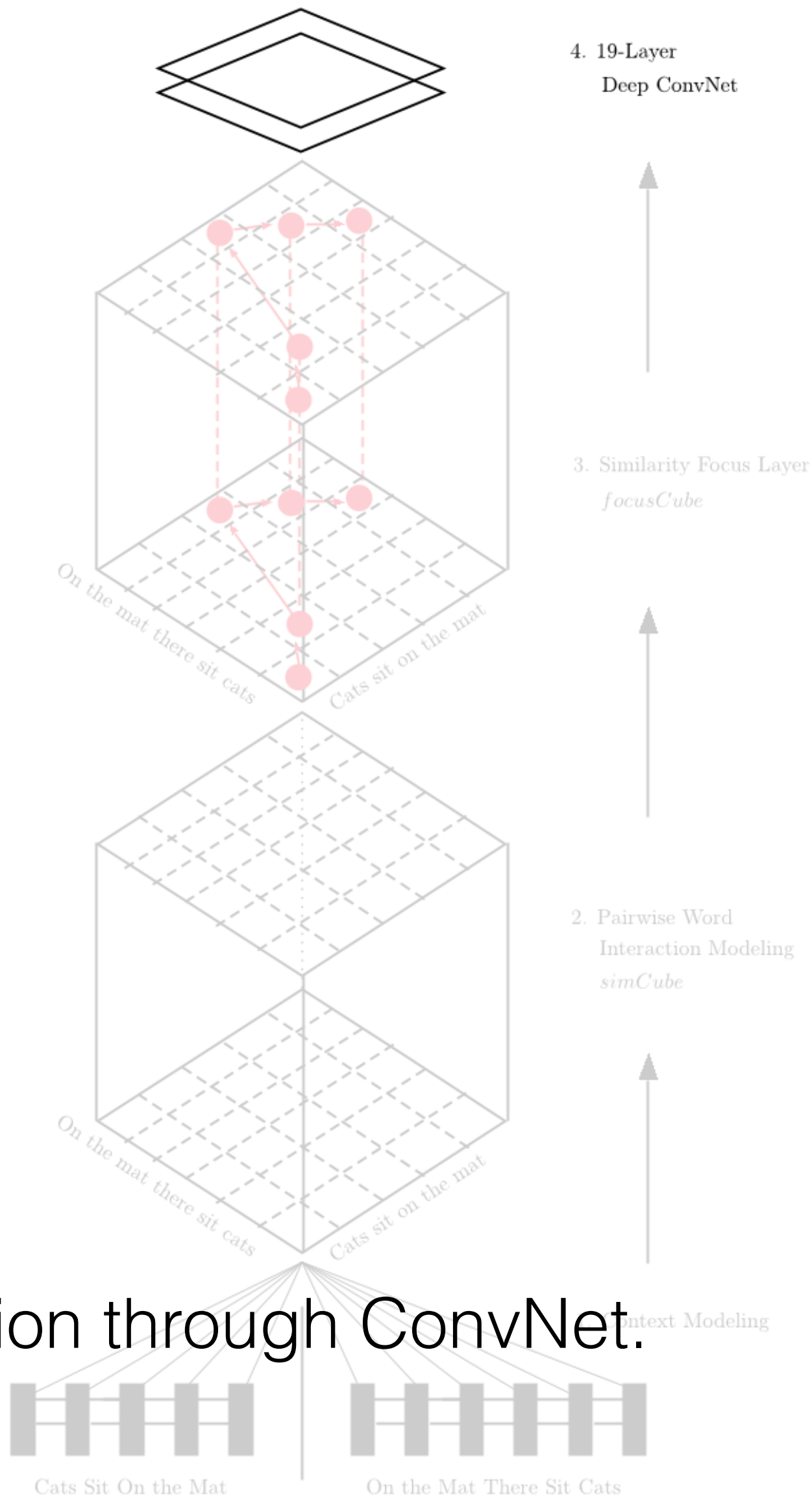
Deep Pairwise Word Model



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	

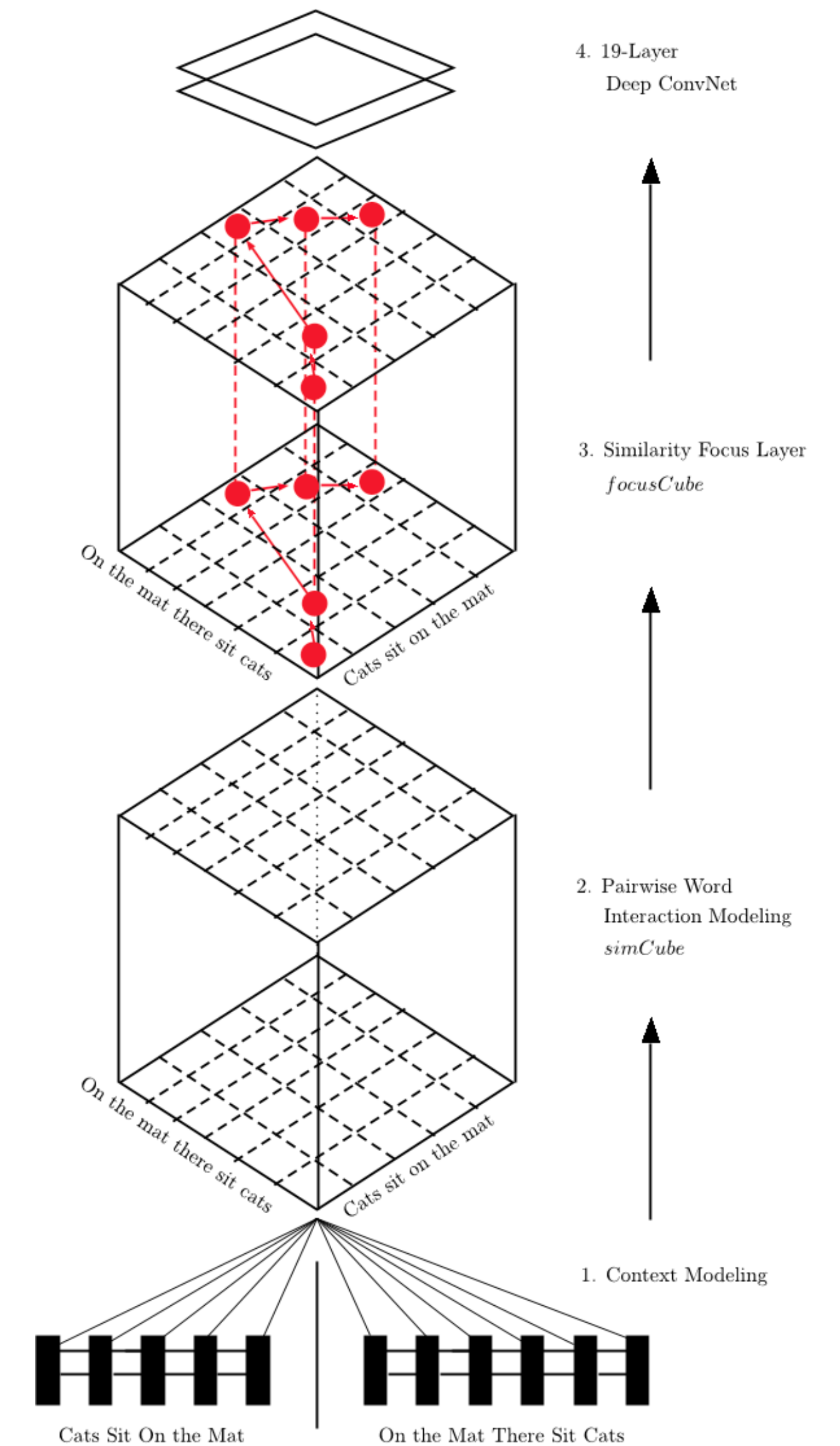
Table 1: Deep ConvNet architecture given two padding size configurations for final classification.



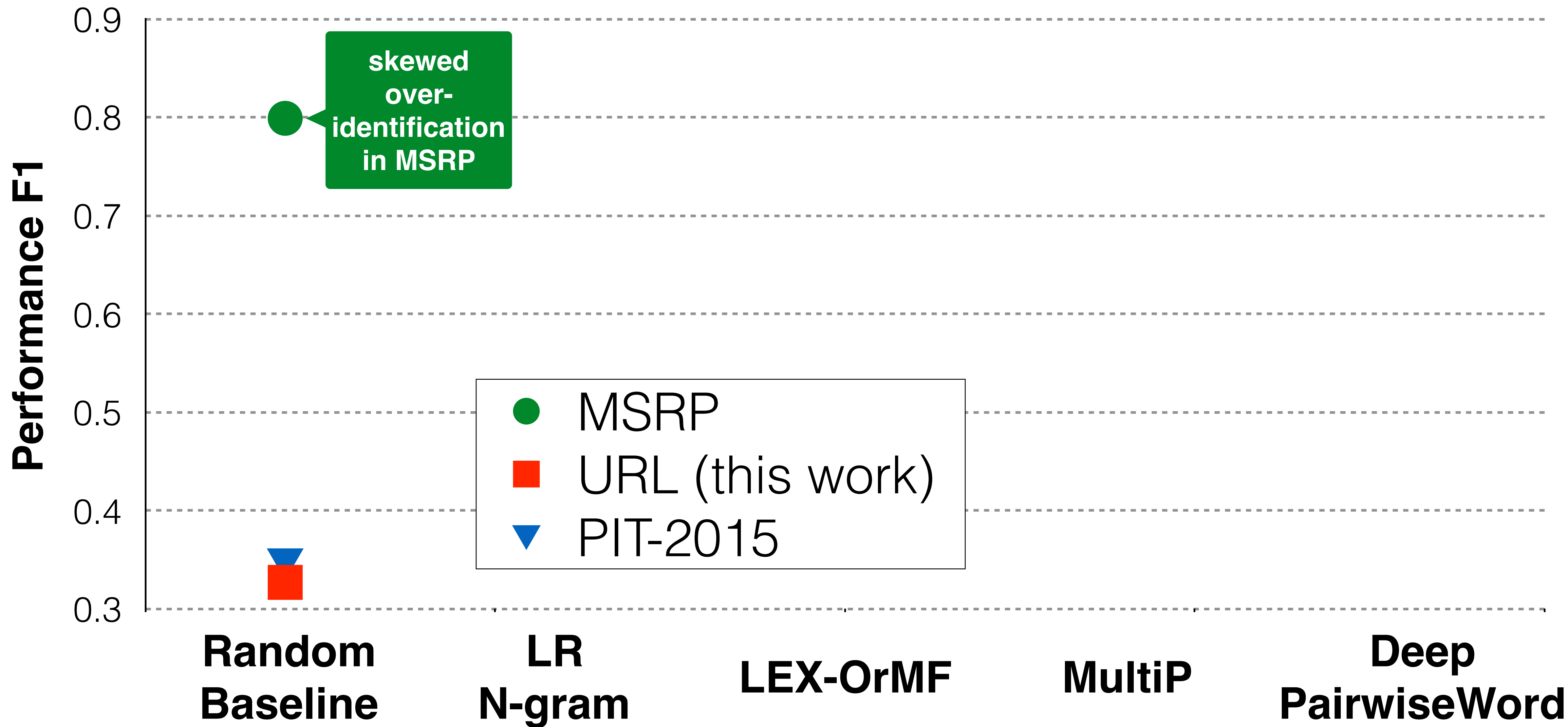
Sentence pair relationship can be identified by pattern recognition through ConvNet.

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**

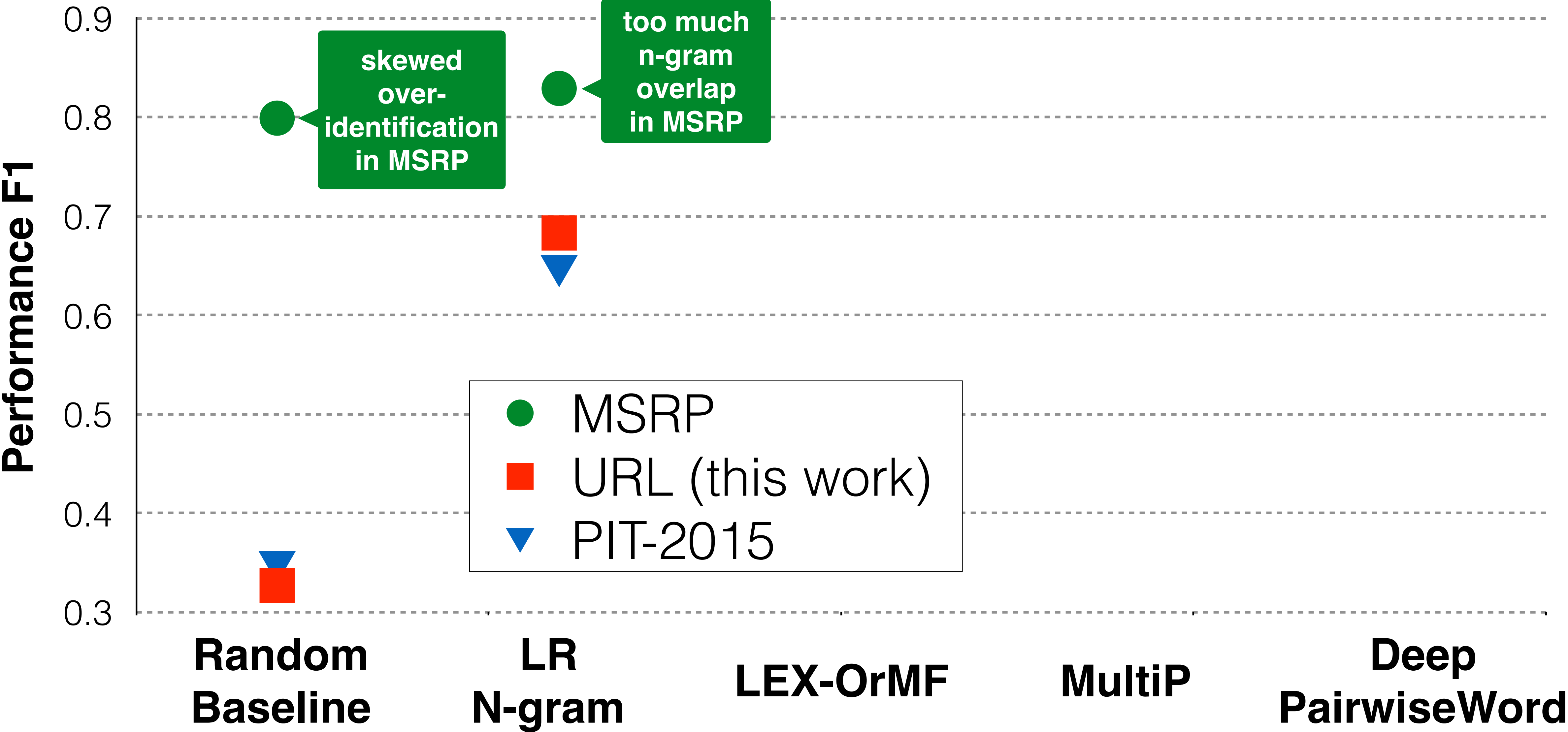


Automatic Paraphrase Identification



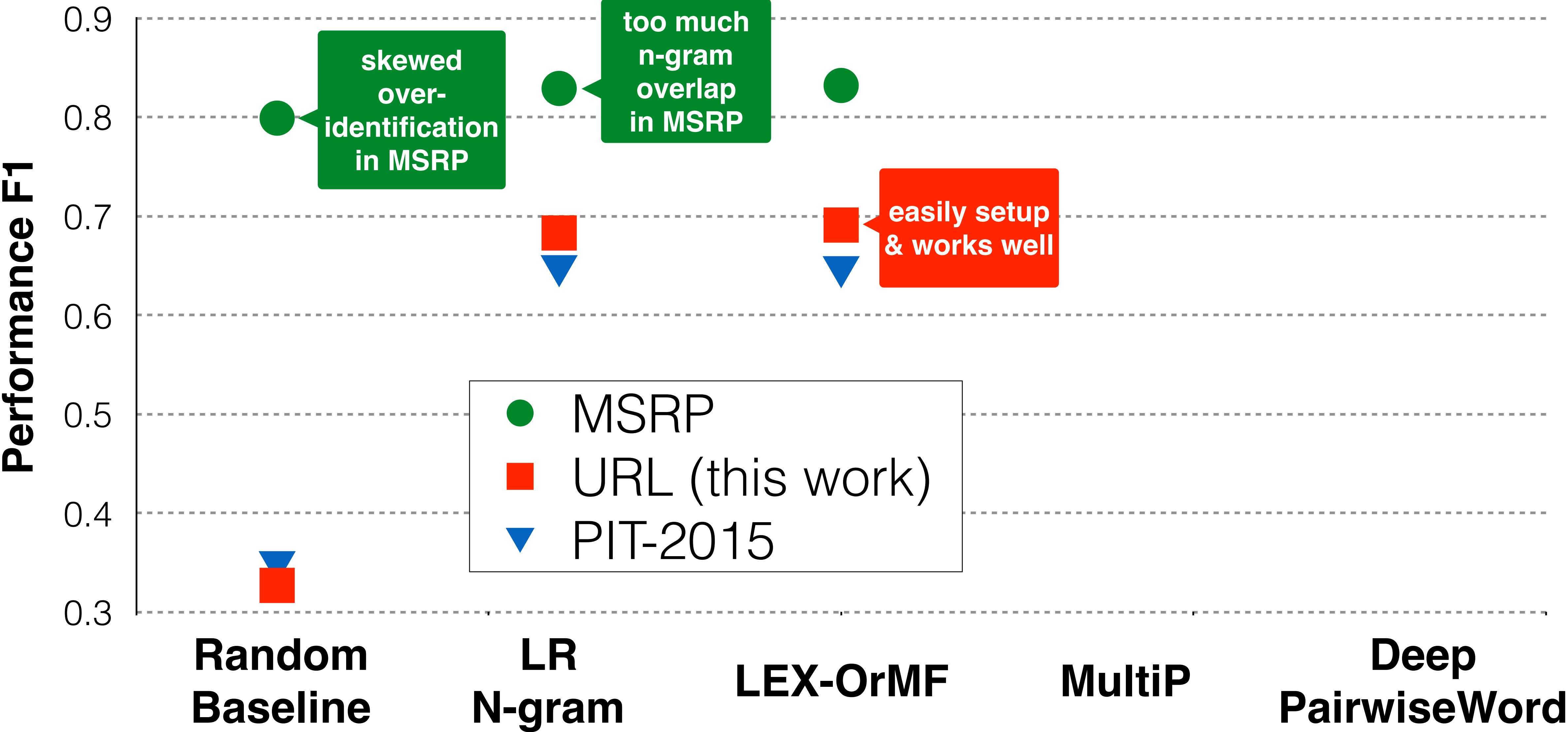
Automatic Paraphrase Identification

MSRP used a SVM classifier
before data annotation



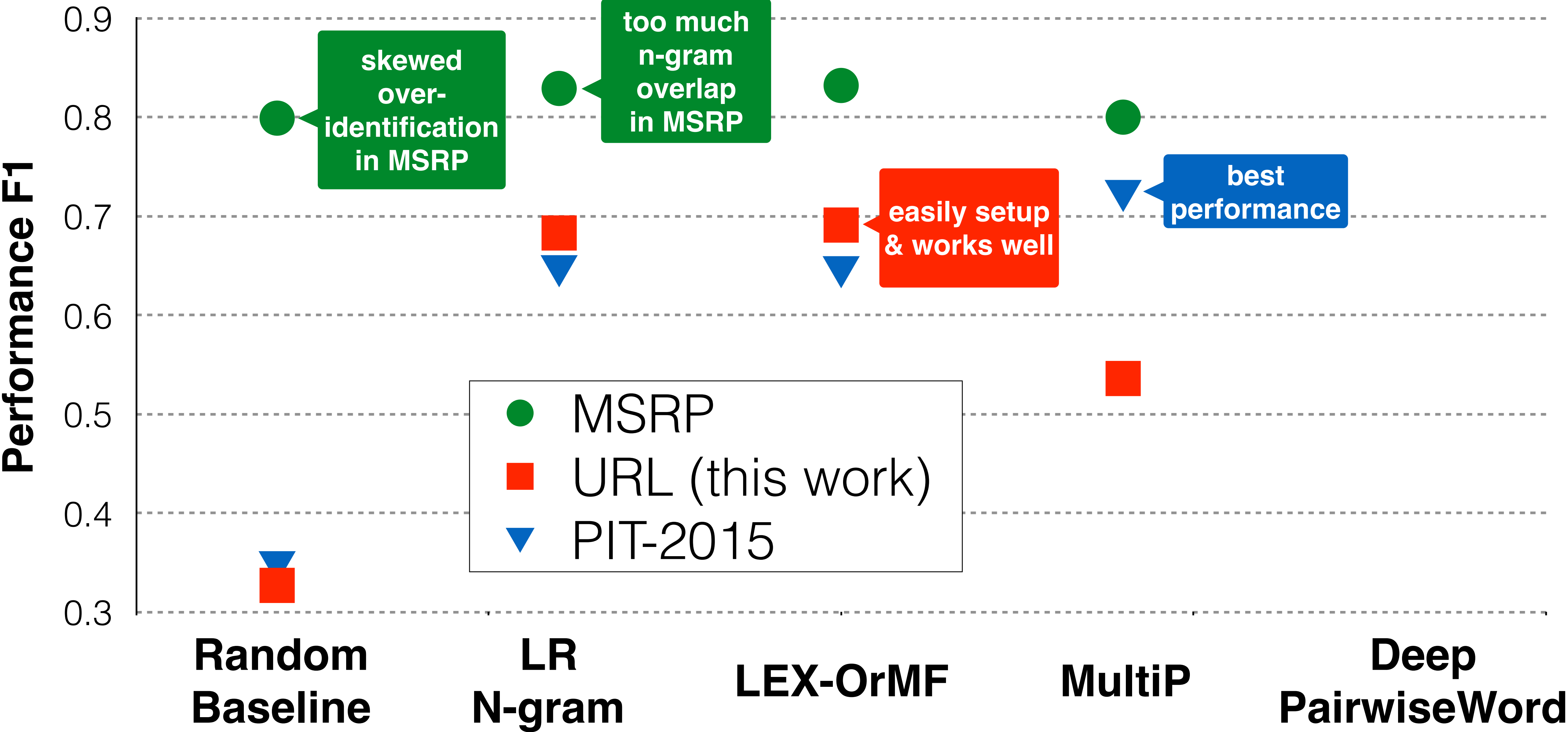
Automatic Paraphrase Identification

MSRP used a SVM classifier
before data annotation



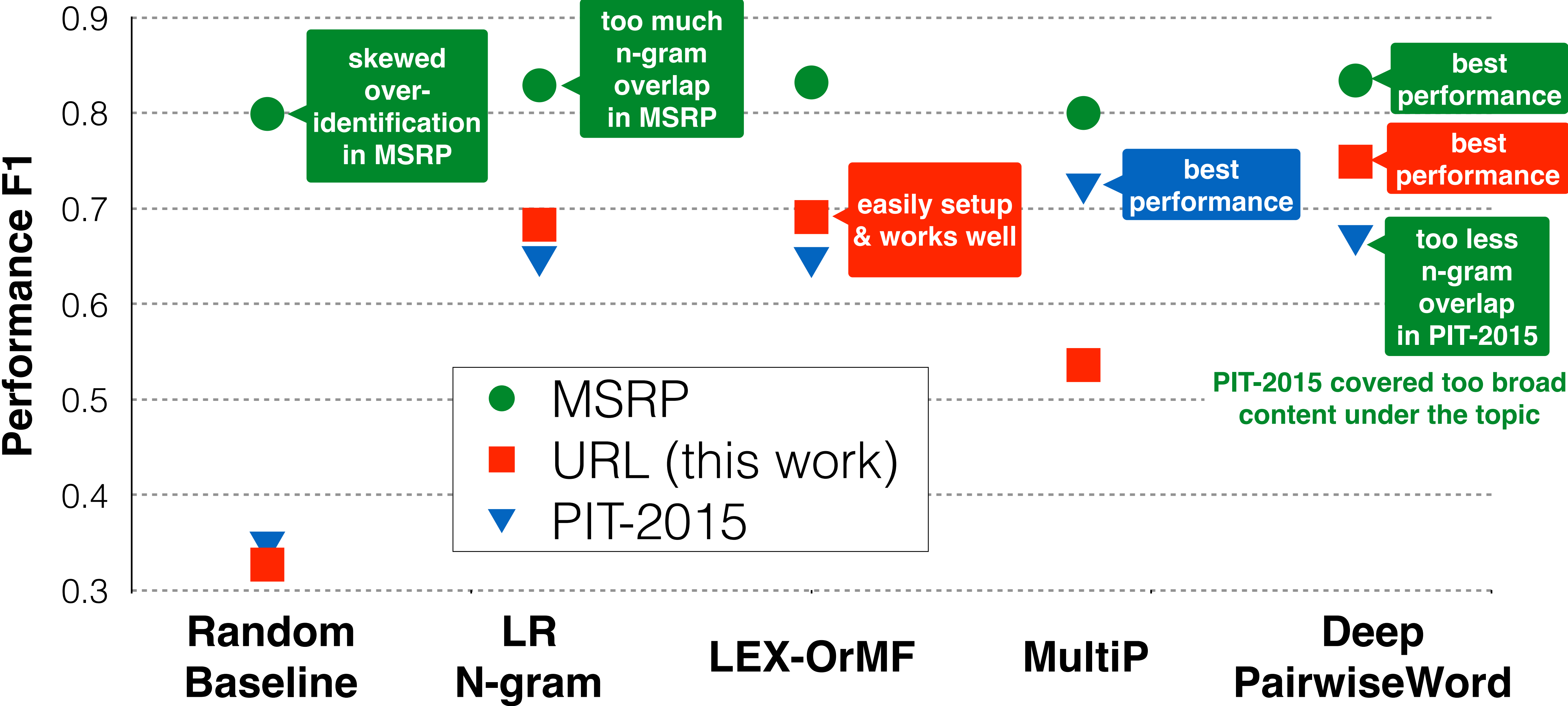
Automatic Paraphrase Identification

MSRP used a SVM classifier
before data annotation

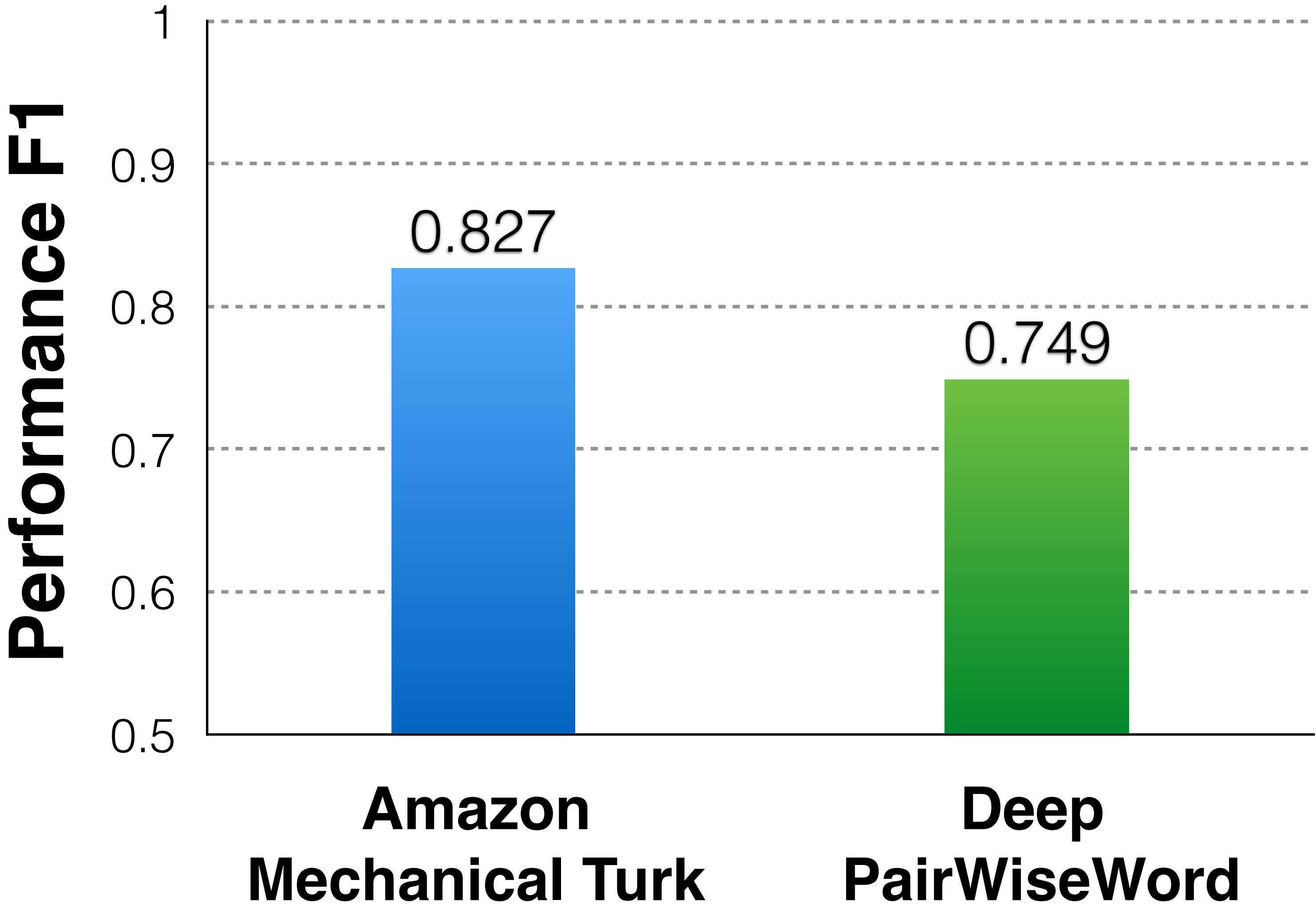


Automatic Paraphrase Identification

MSRP used a SVM classifier
before data annotation



System Performance v.s. Human Upper-bound

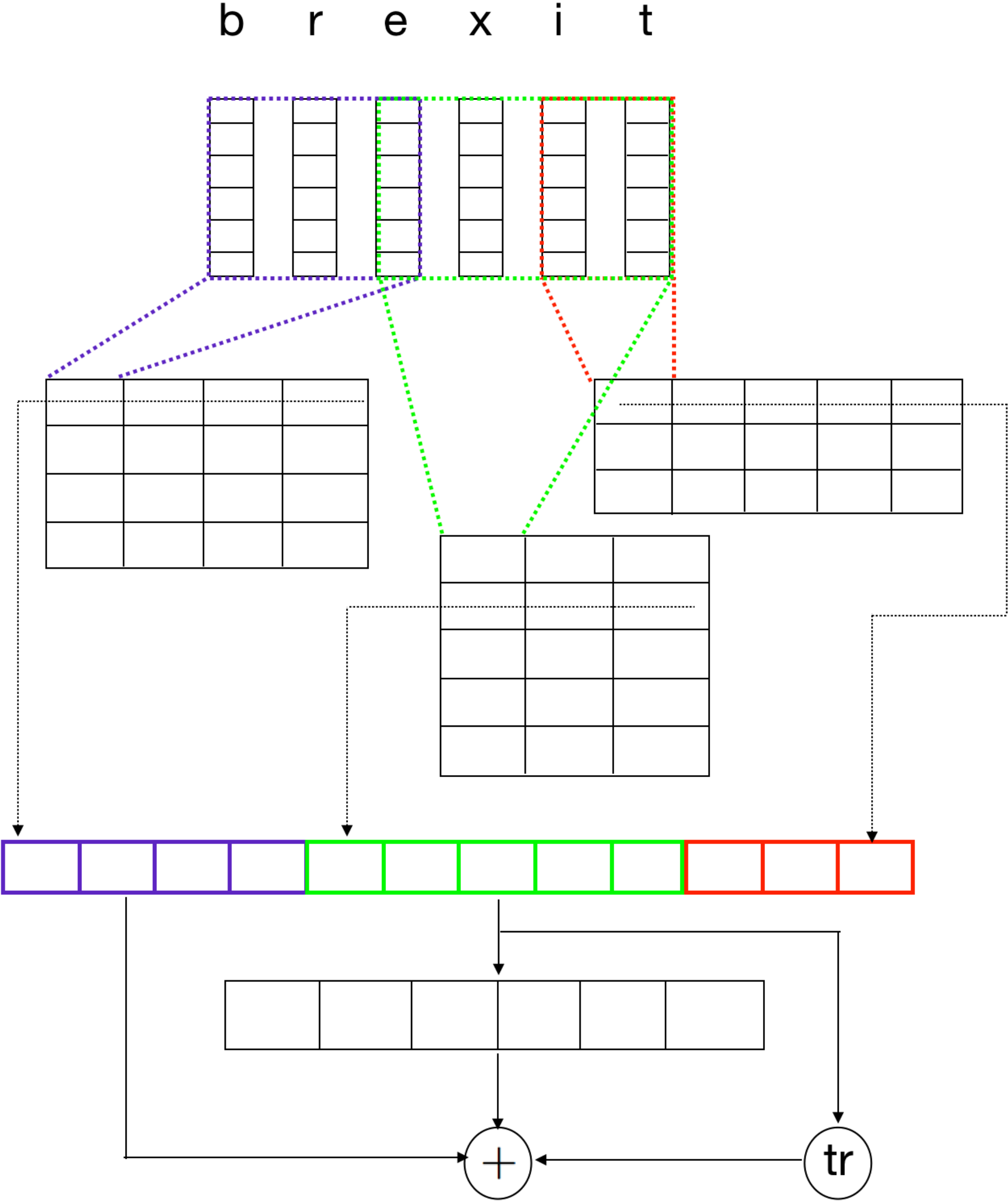


Subword Embedding for Paraphrase Identification

Donald Trump

Donald Trump, DJT, Drumpf, Mr Trump, Idiot Trump, Chump, Evil Donald, #OrangeHitler, Donald @realDonaldTrump, 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, President-elect Trump, President-Elect Trump, President-elect Donald J . Trump, PEOTUS Trump, Emperor Trump

CNN Based Character Embedding



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

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

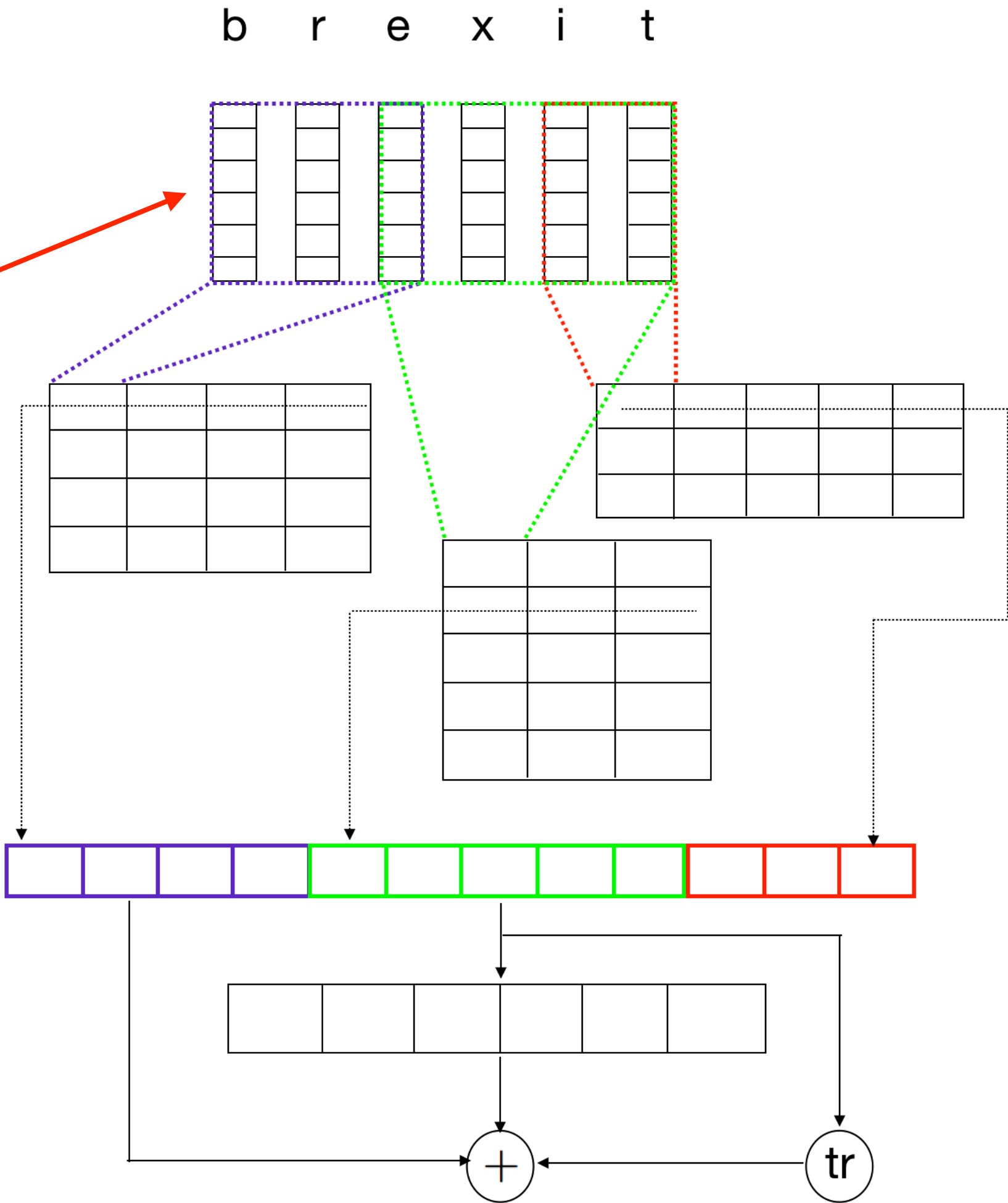
$$\mathbf{z} = \mathbf{t} \odot g(\mathbf{W}_H \mathbf{y} + \mathbf{b}_H) + (\mathbf{1} - \mathbf{t}) \odot \mathbf{y}$$

[1] Kim et al., 2016

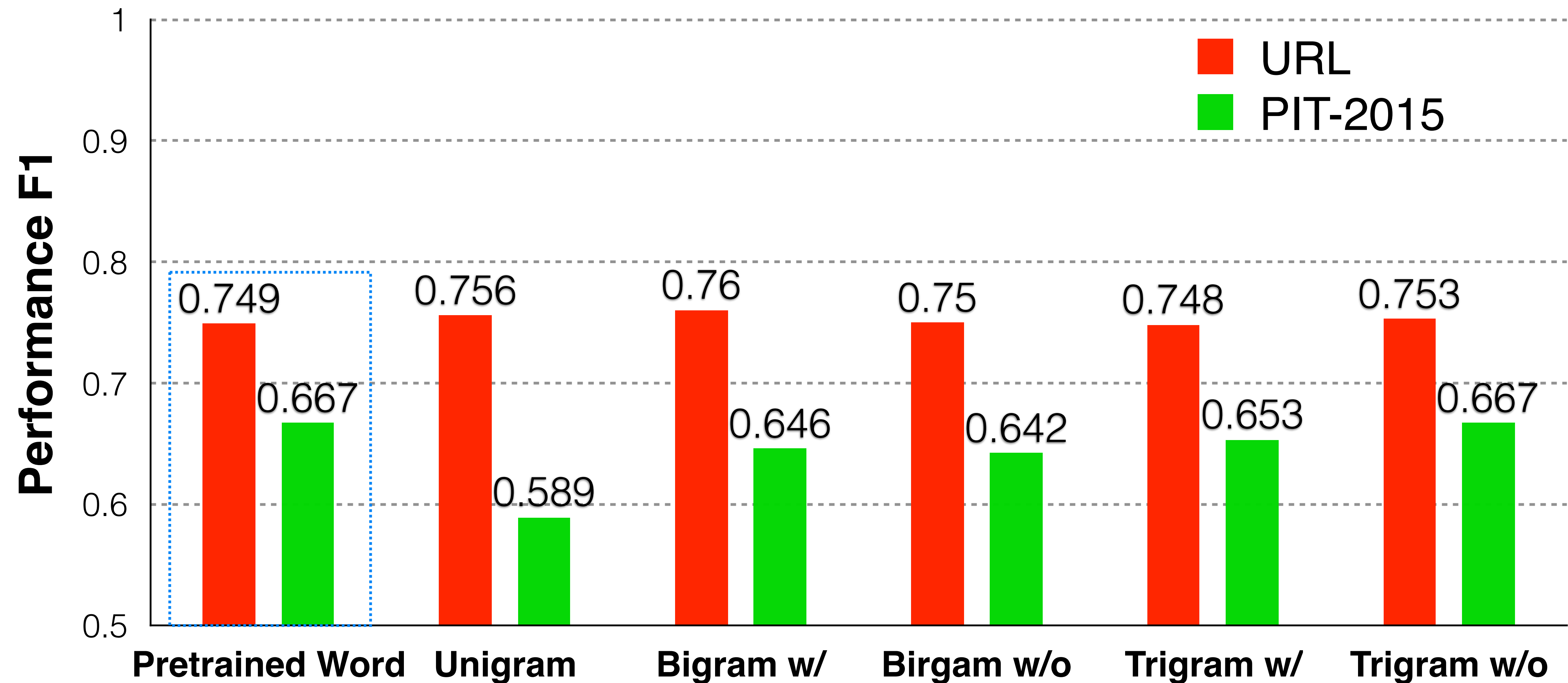
CNN Based Subword Embedding

Unit	Output of $\sigma(\text{brexit})$
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*.



Word Embedding v.s. Subword Embedding



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

