A Continuously Growing Dataset of Sentential Paraphrases

THE OHIO STATE UNIVERSITY

Wuwei Lan[1], Siyu Qiu[2], Hua He[3], and Wei Xu[1]

Penn University of Pennsylvania

UNIVERSITY OF MARYLAND

[1]The Ohio State University, [2]University of Pennsylvania, [3]University of Maryland

Introduction

- Rich application usage of paraphrase
- A simple but powerful URL based data collection method
- Collecting >30k sentential paraphrases with ~70% precision per month
- Largest annotated paraphrase corpus to date

Paraphrase Corpus

Comparison between our data and two existing corpus (MSRP[Dolan et.al 2005] and PIT-2015[Xu et.al 2015]).

Characteristics	MSRP	PIT-2015	Our work
corpora size	6k	14k	51k
+/- balance	✗	✓	✓
manual inspection	✓	✗	✓
over-identification	✗	✓	✓
high precision	✗	✗	✓
dynamic growing	✗	✗	✓

Our Method

Step 1:Collecting tweets that refer to the same URL.

```
for account in {CNN, BBC, New York Times et al.}:
    for tweet in {account.tweets_of_yesterday}:
        if tweet contains URL:
            search_tweet = Twitter.search(key=URL)
            search_tweet.clean()
            if NotOverlap(tweet, search_tweet):
                candidate.append(search_tweet)
    end
end
```

Step 2: Labelling for gold standard corpus

Sampling ~51k candidate sentence pairs
Labeled by Amazon Mechanical Turk

Step 3: Training

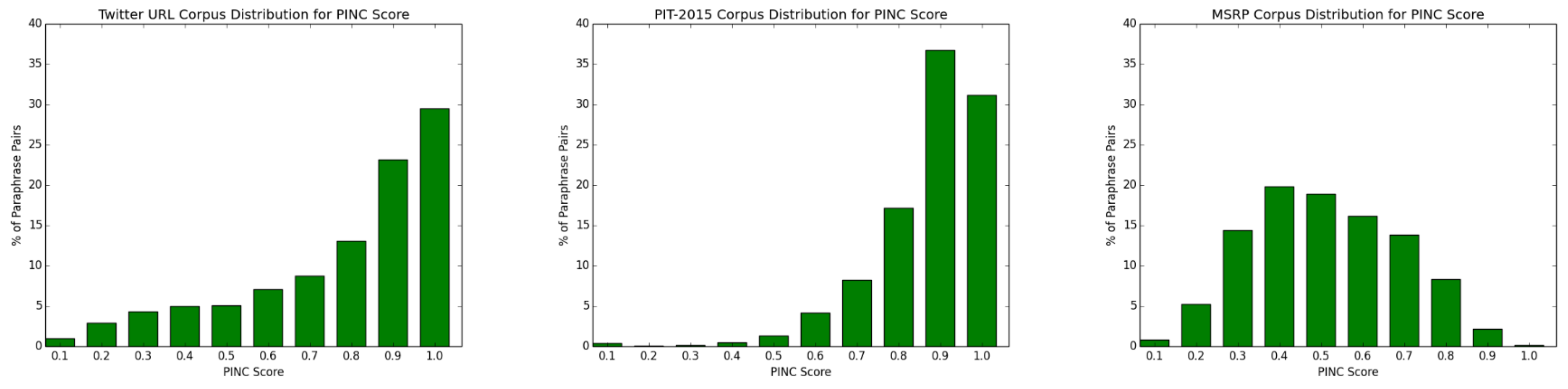
Splitting 42k for training and 9k for testing
Run various paraphrase identification models
Get best classifier M

Step 4: Predicting

M.predict(unlabeled data)

Analysis

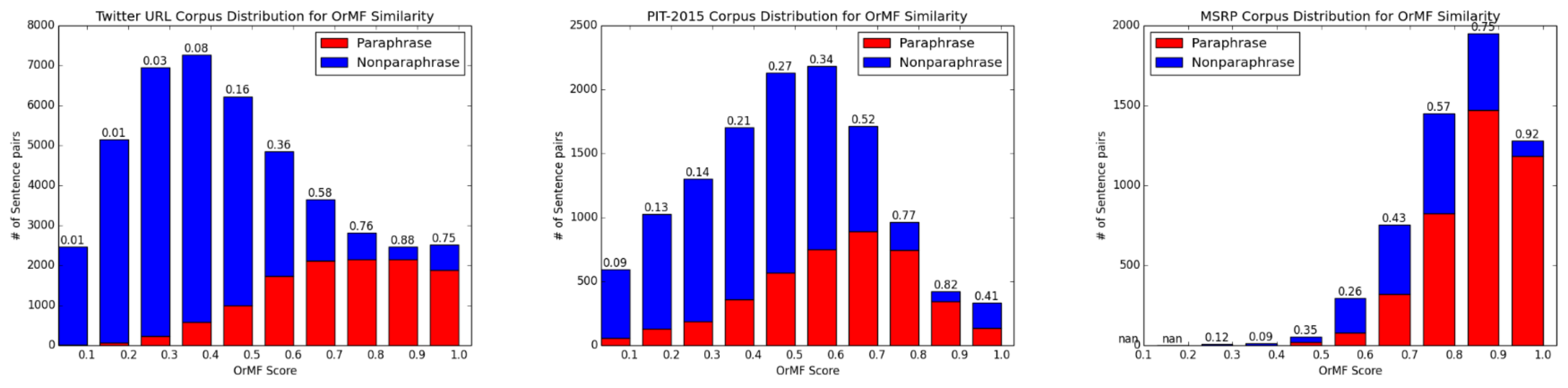
1) Distributional lexical dissimilarity (PINC[Chen et.al 2011])



(a) Twitter URL (b) PIT-2015 (c) MSRP

Lexically divergent in our URL dataset

2) Distributional semantic similarity (OrMF [Guo et.al 2014])



(a) Twitter URL (b) PIT-2015 (c) MSRP

Semantically coherent in our URL dataset

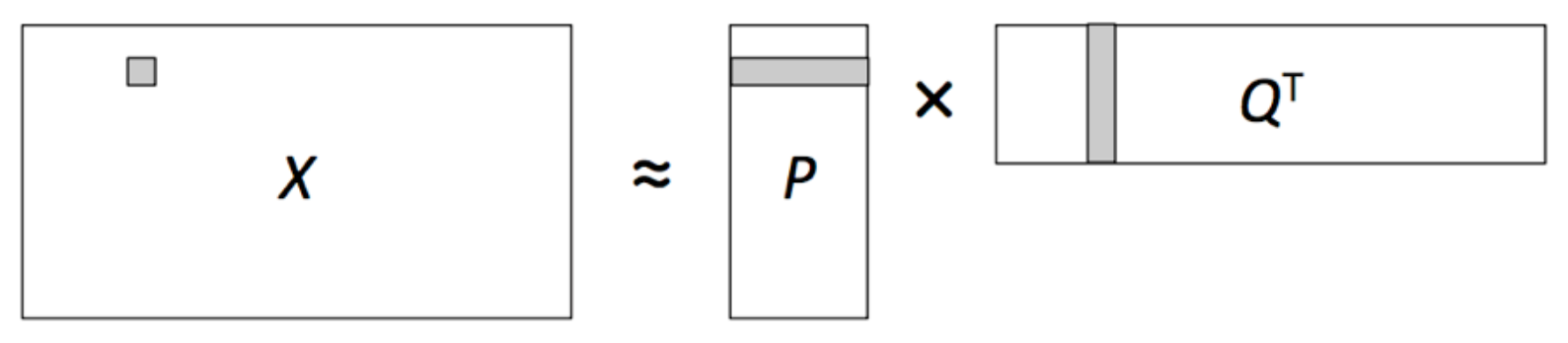
Twitter URL Corpus

Original Tweet	Singer-songwriter Bob Dylan awarded the 2016 Nobel Prize in Literature
Paraphrase	The 2016 Nobel Prize for Literature goes to American singer-songwriter Bob Dylan
	Bob Dylan wins 2016 Nobel Prize for Literature.
	Bob Dylan awarded Nobel Prize for Literature
	The 2016 Nobel Prize in Literature has been awarded to American singer-songwriter Bob Dylan.
Non-Paraphrase	Don’t ask me nothin about nothin . I just might tell you the truth Thank you Mr Dylan .
	The times are a changing . First time in 112 years prize has been awarded to a songwriter .
	Congrats and Respect to the ” Artist’s Artist ” !

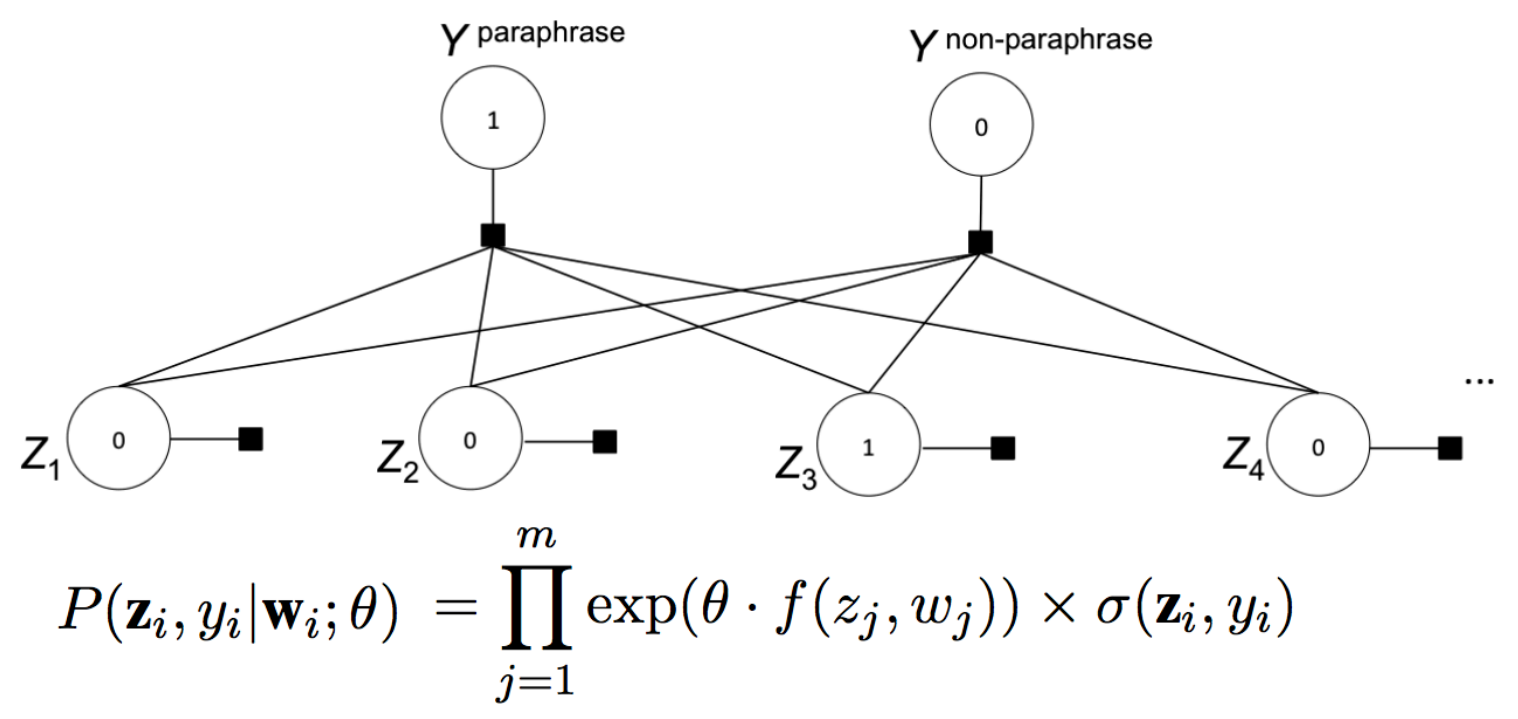
Phrasal Paraphrase

a 15-year-old girl, a 15yr old, a 15 y/o girl
fetuses, fetal tissue, miscarried fetuses
responsible for, guilty of, held liable for, liable for
UVA administrator, UVa official, U-Va. dean, University of Virginia dean
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

Model Comparison

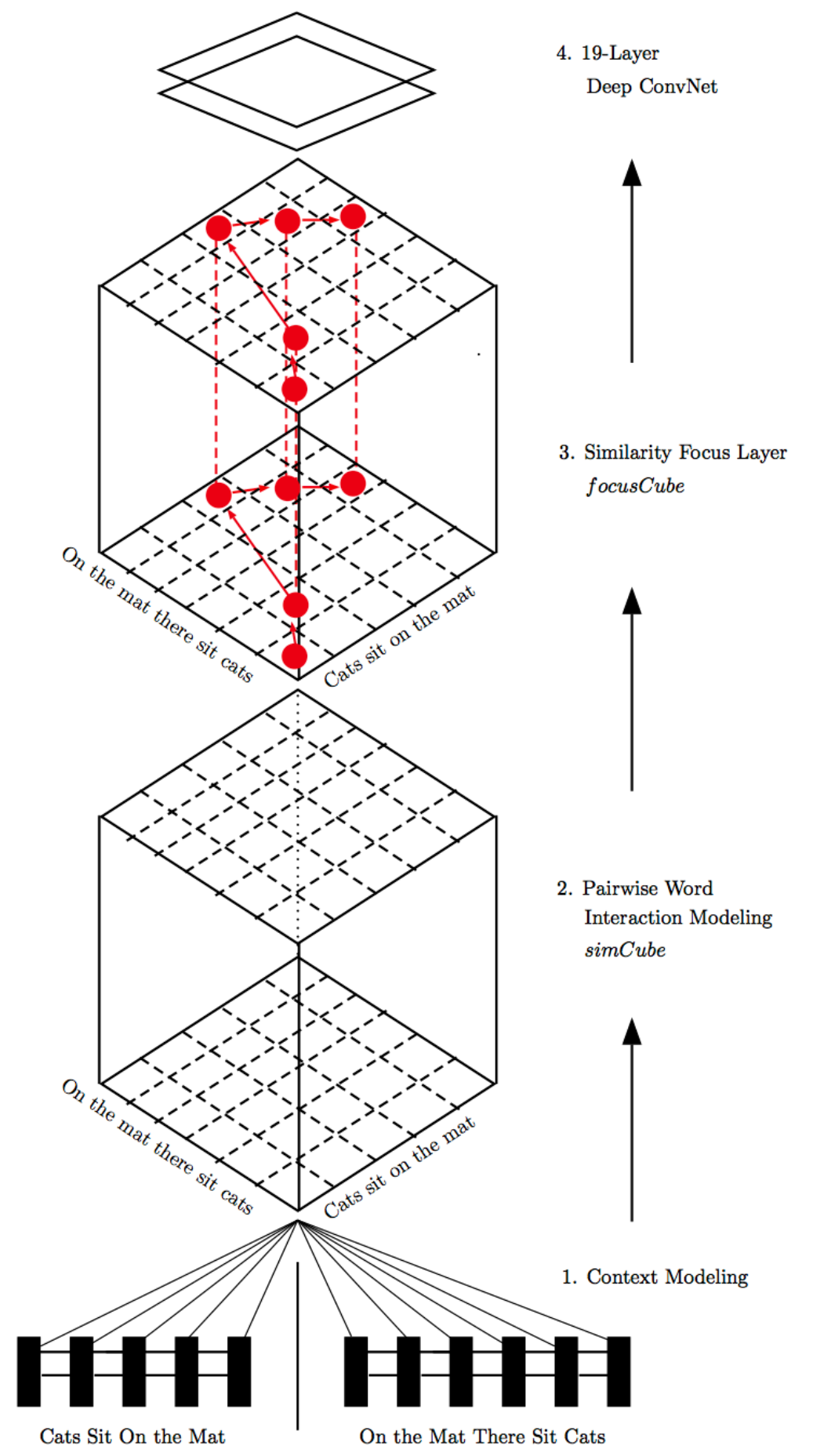


- **LEX-OrMF**[1] (Orthogonal Matrix Factorization[2])

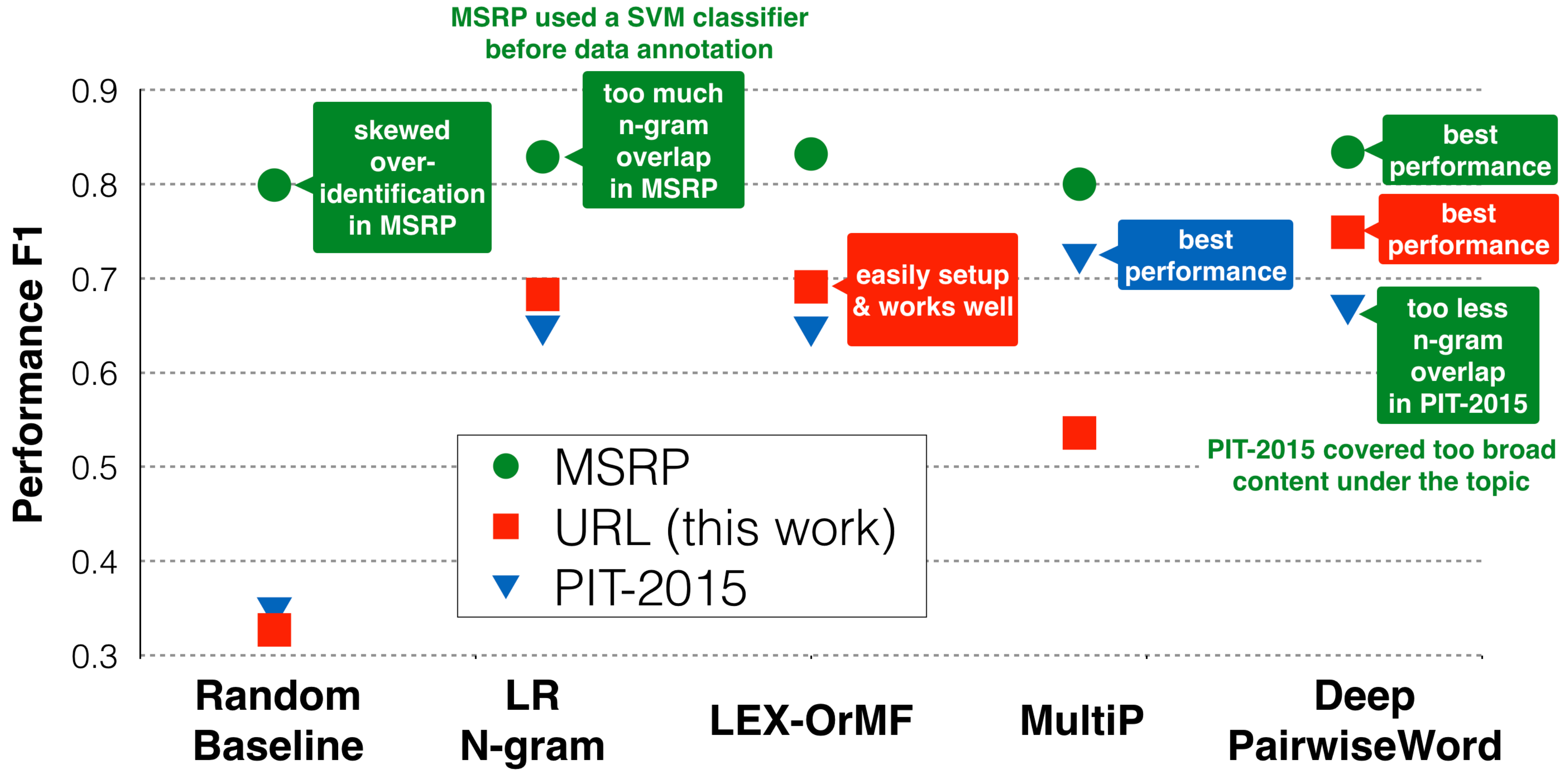


$$P(\mathbf{z}_i, y_i|\mathbf{w}_i;\theta) = \prod_{j=1}^{m} \exp(\theta \cdot f(z_j, w_j)) \times \sigma(\mathbf{z}_i, y_i)$$

- **MultiP**[1] (Multiple Instance Learning)


4. 19-Layer Deep ConvNet
3. Similarity Focus Layer focusCube
2. Pairwise Word Interaction Modeling simCube
1. Context Modeling
On the mat there sit cats
Cats sit on the mat
Cats Sit On the Mat
On the Mat There Sit Cats

- **DeepPairwiseWord**[3] (Deep Neural Networks)



Well balanced in our URL dataset

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

Data and Code

https://github.com/lanwuwei/paraphrase-dataset