Social Media & Text Analysis

lecture 7 - learn large-scale paraphrase from Twitter

Follow @cocoweixu

Instructor: Wei Xu

Website: socialmedia-class.org

Paraphrase

wealthy	**word**	*rich*
the king's speech	**phrase**	*His Majesty's address*
… the forced resignation of the CEO of Boeing, Harry Stonecipher, for …	**sentence**	*… after Boeing Co. Chief Executive Harry Stonecipher was ousted from …*

Application

Information Extraction

end_job (Harry Stonecipher, Boeing)

↑ **extract**

… the forced resignation of the CEO of Boeing, Harry Stonecipher, for …

… after Boeing Co. Chief Executive Harry Stonecipher was ousted from …

Wei Xu, Raphael Hoffmann, Le Zhao, Ralph Grishman.
"Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction" In ACL (2013)

Application

Question Answering

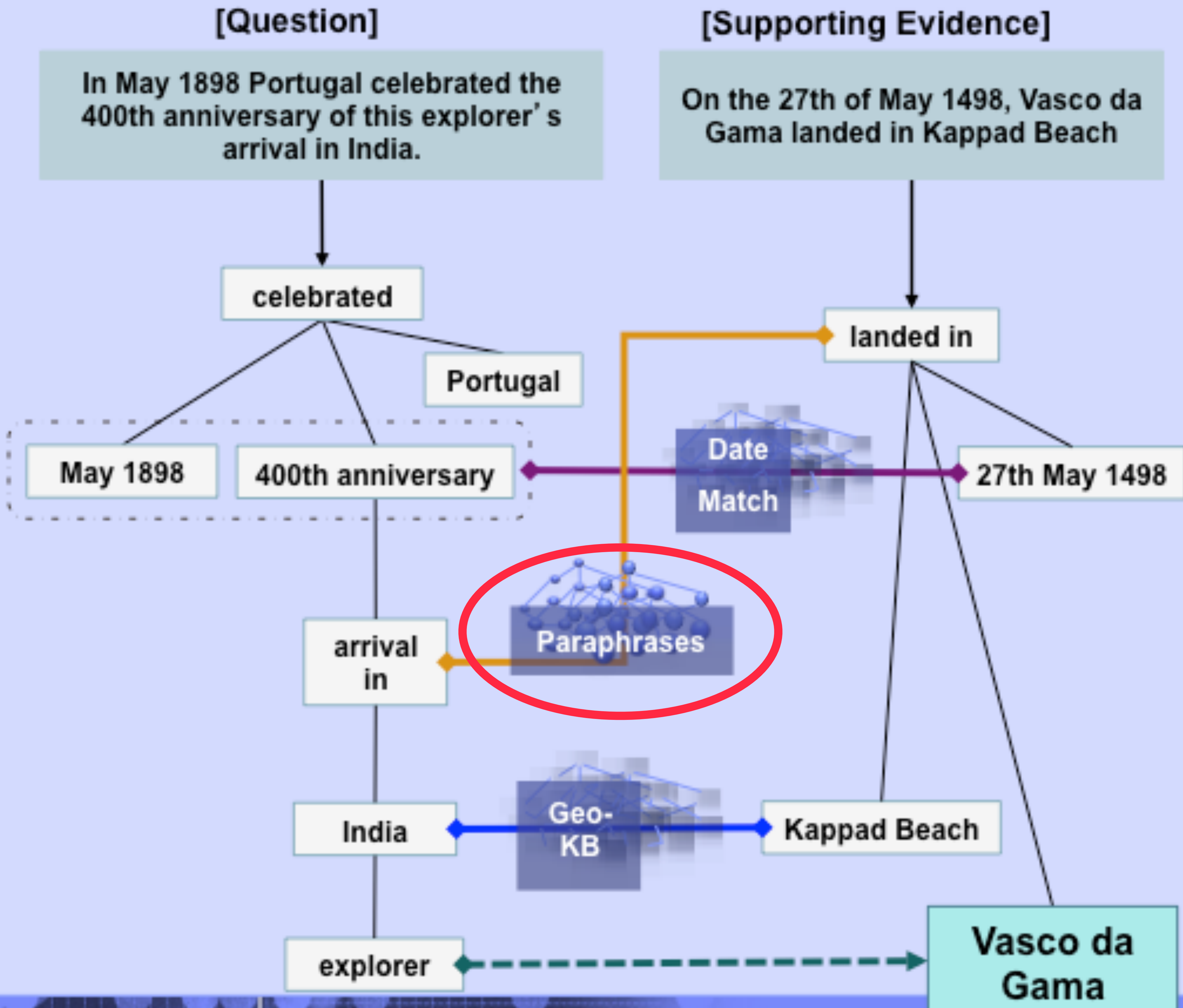
Who is the CEO stepping down from Boeing?

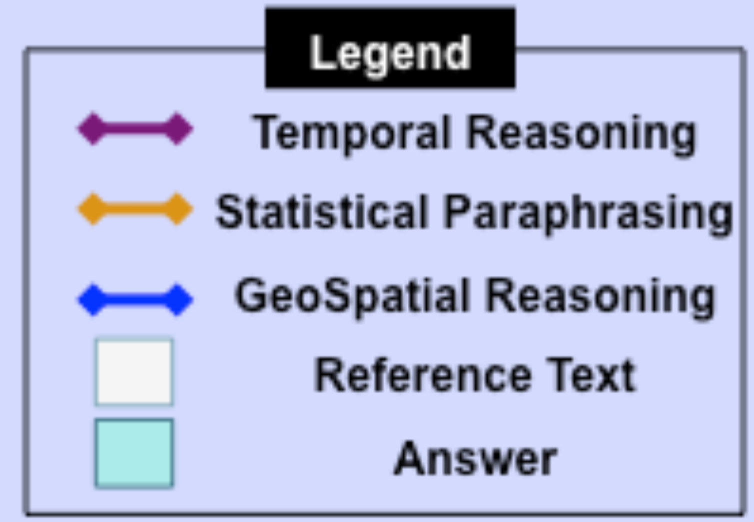
match

… the forced resignation of the CEO of Boeing, Harry Stonecipher, for …

… after Boeing Co. Chief Executive Harry Stonecipher was ousted from …

Watson leverages multiple algorithms to perform deeper analysis





Stronger evidence can be much harder to find and score…

- Search far and wide
- Explore many hypotheses
- Find judge evidence
- Many inference algorithms

(courtesy:Salim Roukos)

Application

Text Simplification

They are culturally akin to the coastal peoples of Papua New Guinea.

↓

Their culture is like that of the coastal peoples of Papua New Guinea.

NSF EAGER: “Simplification as Machine Translation” (2014 ~ 2015)

Wei Xu, Chris Callison-Burch. “Problems in Current Text Simplification Research: New Data Can Help” TACL (2015)

Application

Stylistic Rewriting

Palpatine:
If you will not be turned, you will be destroyed!

If you will not be turn'd, you will be undone!

Luke:
Father, please! Help me!

Father, I pray you! Help me!

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry. "Paraphrasing for Style" In COLING (2012)

Paraphrase Data

80s
WordNet

'01
Novels
Barzilay
McKeown

'04
News
Dolan
Quirk
Brockett

'05 '13
Bi-Text
Callison-Burch
Ganitkevitch
Van Durme
Bernard

'11
Video
Chen
Dolan

'12
Shakespeare
Xu
Ritter
Dolan
Grishman
Cherry

(This Work)
Twitter
Xu
Ritter
Callison-Burch
Dolan
Ji

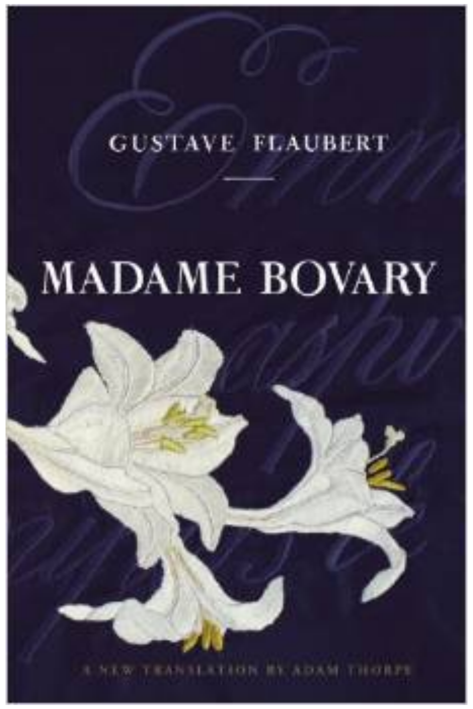
Paraphrase Research

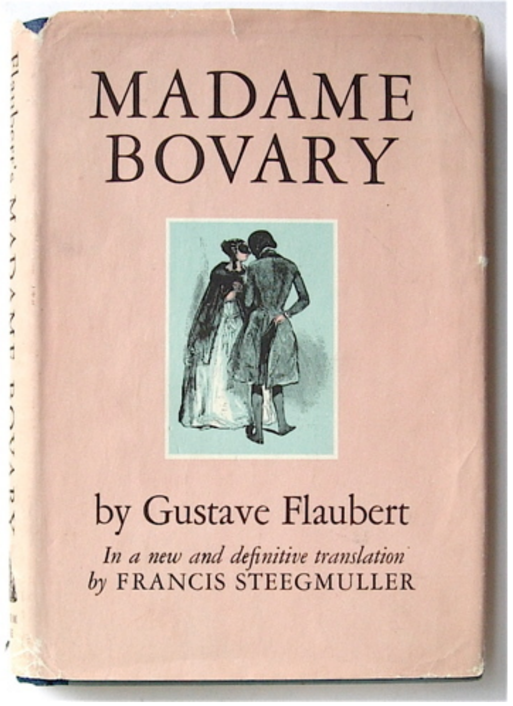
WordNet **Novel** **News** **Bi-Text** **Video** **Shakespeare** **Twitter**

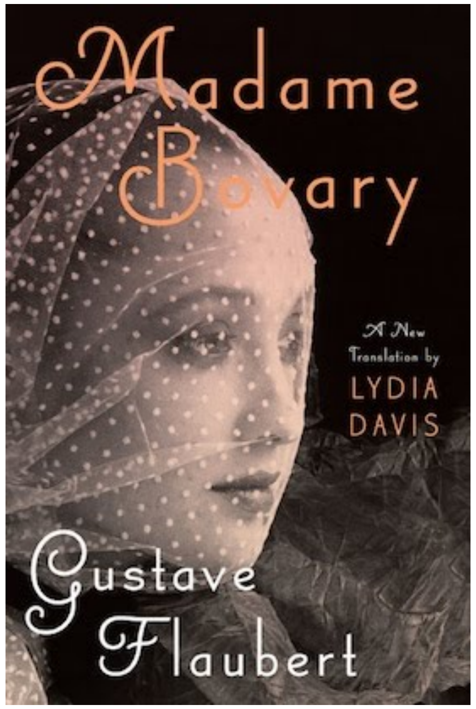
information extraction
question answering
semantic similarity
semantic parsing
text-to-text generation
…

but, primarily for formal language usage and well-edited text

Previous Work







multiple English translations of novels

(Barzilay and McKeown, 2001)

Previous Work

only a few hundreds news agencies
report big events
using formal language

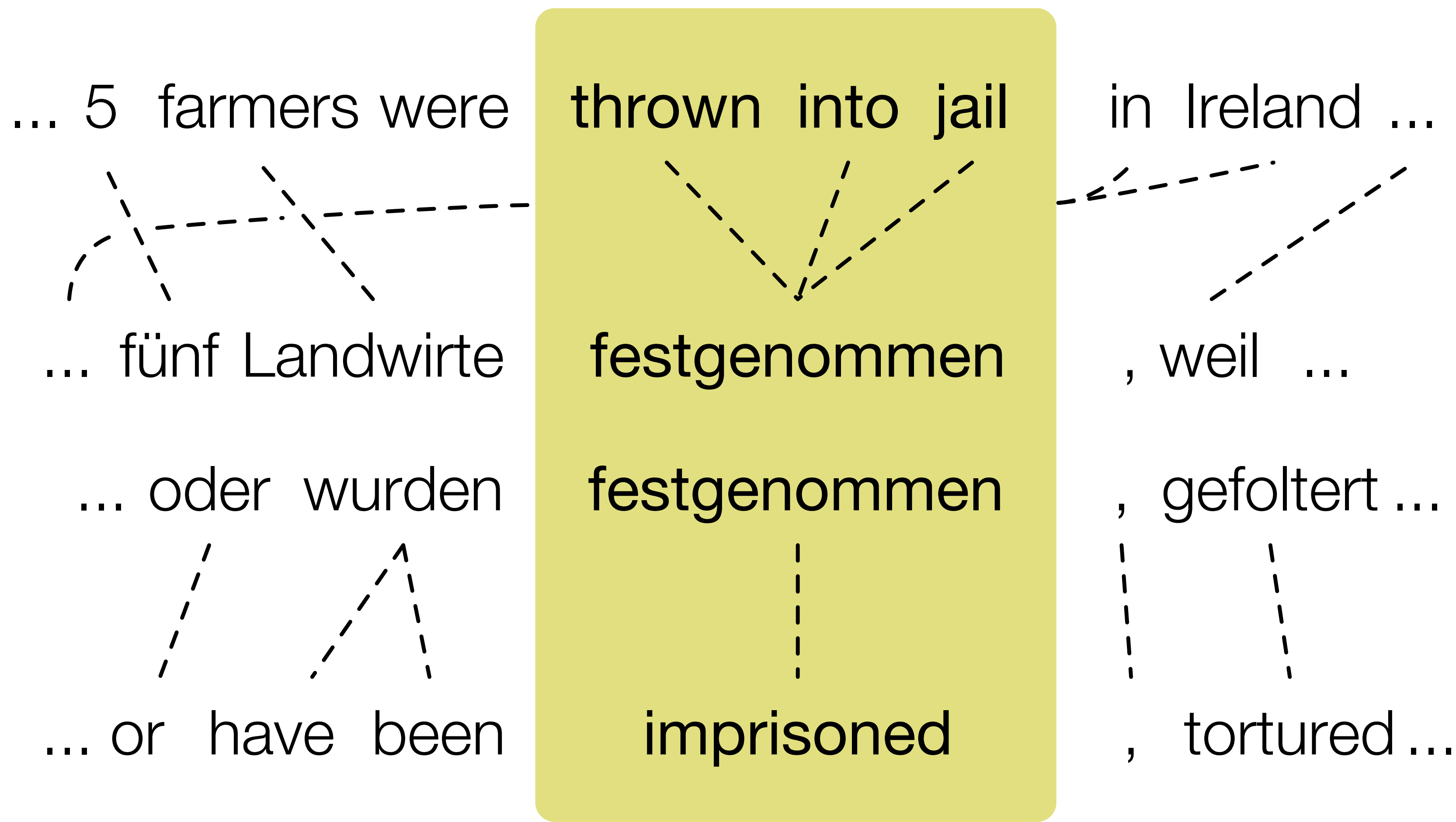
(Dolan, Quirk and Brockett, 2004; Dolan and Brockett, 2005; Brockett and Dolan, 2005)

Previous Work

ask dozens of annotators to write
one sentence for a short video (<= 10 seconds)

(Chen and Dolan, 2011)

Previous Work



pivoting through bilingual text
from European Parliament proceedings,
multilingual websites etc.

(Bannard and Callison-Burch, 2005; Ganitkevitch, Van Durme and Callison-Burch, 2013)

Twitter as a new resource

Wei Xu, Alan Ritter, Ralph Grishman.
"A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2014)

Twitter as a powerful resource

thousands of users
talk about both big and micro events
using formal, informal, erroneous language

Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji.
"Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Enables new applications

Noisy Text Normalization

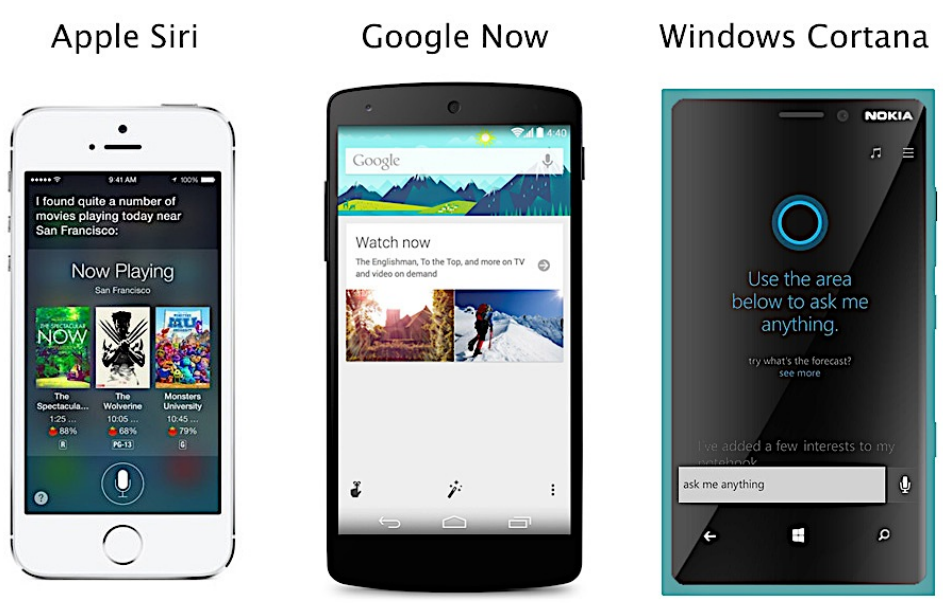
oscar nom'd doc → *Oscar-nominated documentary*

don't want for → *don't wait for*

Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao. "Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-gram Models" In EMNLP (2011)

Enables new applications

Human-computer Interaction



who wants to get a beer?	→

want to get a beer?
who else wants to get a beer?
who wants to go get a beer?
who wants to buy a beer?
who else wants to get a beer?
trying to get a beer?

… (21 different ways)

Wei Xu, Alan Ritter, Ralph Grishman.
"Gathering and Generating Paraphrases from Twitter with Application to Normalization" In BUCC (2013)

Enables new applications

Listen & Speak
Like a Native Speaker

Language Education

Aaaaaaaaand stephen curry <u>is on fire</u>

↕

<u>What a incredible performance</u> from Stephen Curry

(Ongoing work with Quanze Chen and Chris Callison-Burch)

Enables new applications

Sentiment Analysis 🙂 **or** 🙁 **?**

This nets vs bulls game is great

This Nets vs Bulls game is nuts

Wowsers to this nets bulls game

this Nets vs Bulls game is too live

This Nets and Bulls game is a good game

This netsbulls game is too good

This NetsBulls series is intense

Learn Paraphrases

Learn Paraphrases

identify parallel sentences automatically from Twitter's big data stream

Mancini has been sacked by Manchester City
Mancini gets the boot from Man City

Yes!

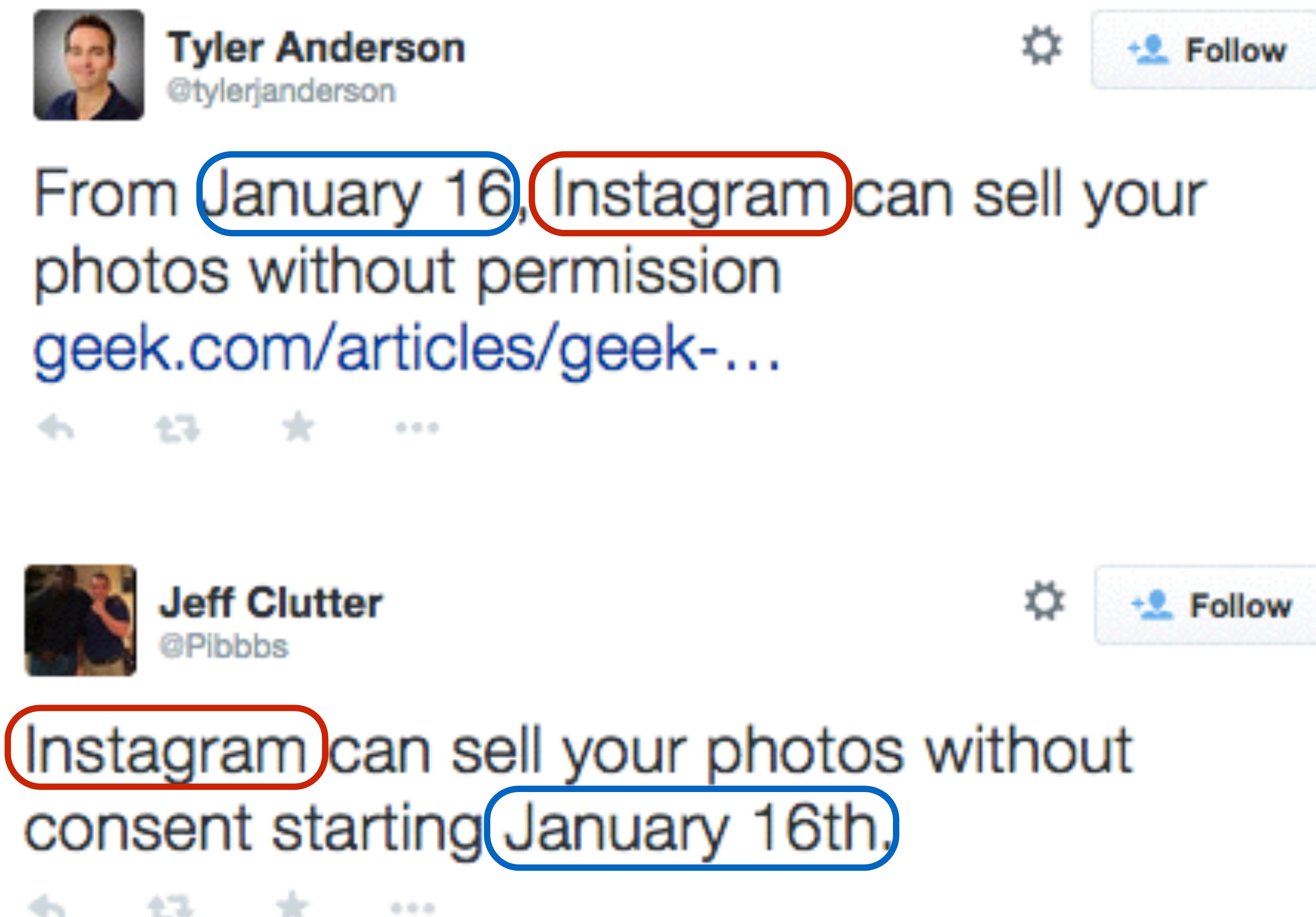
WORLD OF JENKS IS ON AT 11
World of Jenks is my favorite show on tv

No!

Early Attempts on Twitter

- 1242 tweet pairs, tracking celebrity & hashtags (Zanzotto, Pennacchiotti, Tsioutsiouliklis, 2011)
- named entity + date (Xu, Ritter, Grishman, 2013)
- bilingual posts (Ling, Dyer, Black, Trancoso, 2013)

Named Entity + Time



Wei Xu, Alan Ritter, Ralph Grishman.
“Gathering and Generating Paraphrases from Twitter with Application to Normalization” In BUCC (2013)

Self-translation

(Ling, Dyer, Black and Trancoso, 2013)

Design a Model

Train it on data

A Challenge

<u>Mancini</u> has been sacked by Manchester City

<u>Mancini</u> gets the boot from Man City

very short
lexically divergent

(less word overlap, even in high-dimensional space)

Previous Methods

Algorithm	Reference	Description	Supervision
Vector Based	Mihalcea et al. (2006)	cosine similarity with tf-idf weighting	unsupervised
ESA	Hassan (2011)	explicit semantic space	unsupervised
KM	Kozareva and Montoyo (2006)	combination of lexical and semantic features	supervised
LSA	Hassan (2011)	latent semantic space	unsupervised
RMLMG	Rus et al. (2008)	graph subsumption	unsupervised
MCS	Mihalcea et al. (2006)	combination of several word similarity measures	unsupervised
WTMF	Guo and Diab (2012)	latent semantics model of missing words	unsupervised
STS	Islam and Inkpen (2007)	combination of semantic and string similarity	unsupervised
SSA	Hassan (2011)	salient semantic space	unsupervised
QKC	Qiu et al. (2006)	sentence dissimilarity classification	supervised
ParaDetect	Zia and Wasif (2012)	PI using semantic heuristic features	supervised
SDS	Blacoe and Lapata (2012)	simple distributional semantic space	supervised
matrixJcn	Fernando and Stevenson (2008)	JCN WordNet similarity with matrix	unsupervised
FHS	Finch et al. (2005)	combination of MT evaluation measures as features	supervised
PE	Das and Smith (2009)	product of experts	supervised
WDDP	Wan et al. (2006)	dependency-based features	supervised
SHPNM	Socher et al. (2011)	recursive autoencoder with dynamic pooling	supervised
MTMETRICS	Madnani et al. (2012)	combination of eight machine translation metrics	supervised
LEXLATENT	Ji and Eienstein (2013)	combination of latent space and lexical features	supervised

mostly based on sentence similarity of surface words or latent semantics

Design a Model

At-least-one-anchor Assumption

two sentences about the same topic are paraphrases
if and only if
they contain at least one word pair that is a paraphrase **anchor**

That boy Brook Lopez with a deep ***3***
brook lopez hit a ***3***

Yes!

Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji.
"Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Another Challenge

not every word pair of similar meaning indicates sentence-level paraphrase

<u>Iron Man</u> ***3*** *was brilliant fun*
<u>Iron Man</u> ***3*** *tonight see what this is like*

No!

Solution:

a discriminative model using features at word-level

Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji.
"Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Features

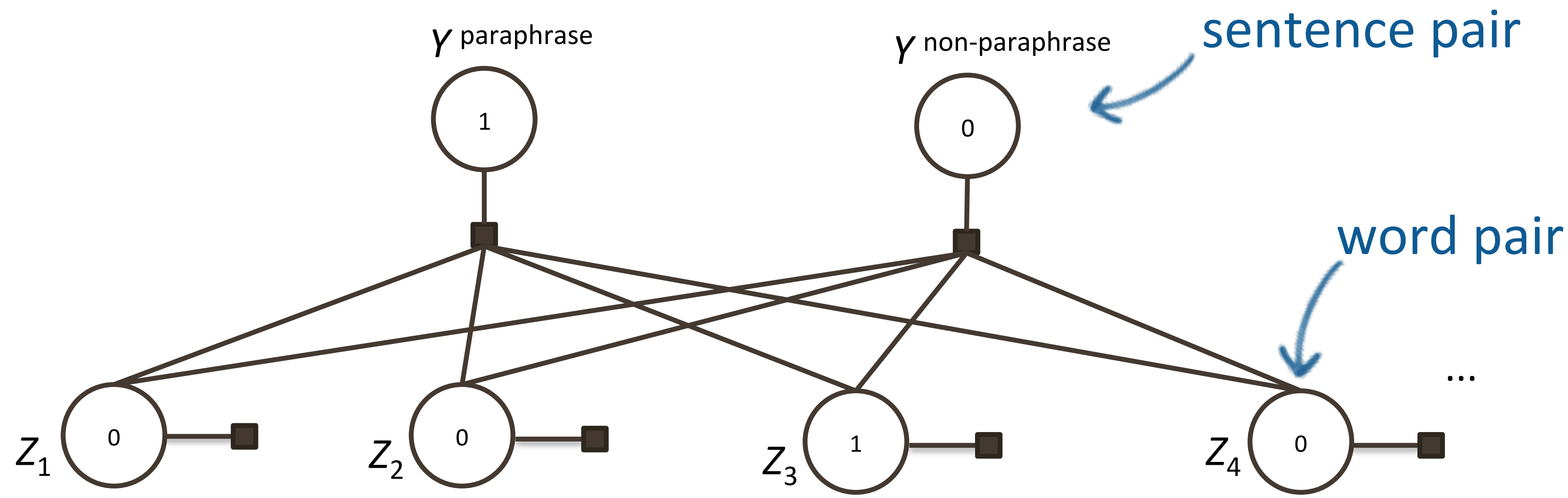
- String Features:
  - words, stemmed forms, normalized forms
  - same, similar or dissimilar
- POS Features:
  - fine grained tags:
    “a”, “be”, “do”, “have”,
    “get”, “go”, “follow”, “please”
- Topical Features:
  - word significantly associated with each topic
  - e.g. “3” for basketball; “RIP” for death events

Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. “Extracting Lexically Divergent Paraphrases from Twitter” In TACL (2014)

Multi-instance Learning Paraphrase Model

Manti bout to be the ***next*** *<u>Junior Seau</u>*

Teo is the little ***new*** *<u>Junior Seau</u>*



manti | teo

diff_word
same_pos_nn
both_sig
…

be | is

same_stem
same_pos_be
not_both_sig
…

next | new

diff_word
same_pos_jj
both_sig
…

manti | little …

diff_word
diff_pos_nn
diff_pos_jj
not_both_sig
…

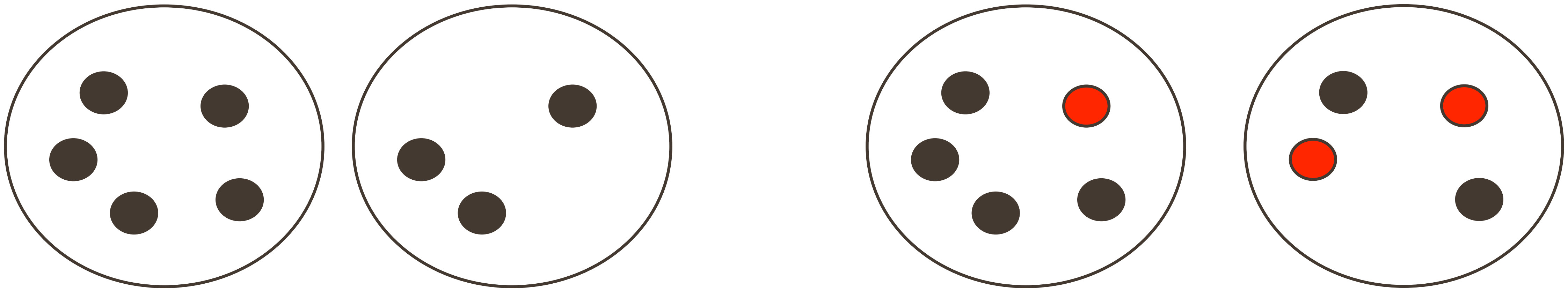
features

Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji.
"Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

[Mini Tutorial]
Multi-instance Learning

Instead of labels on each individual instance,
the learner only observes labels on bags of instances.

Negative Bags | Positive Bags



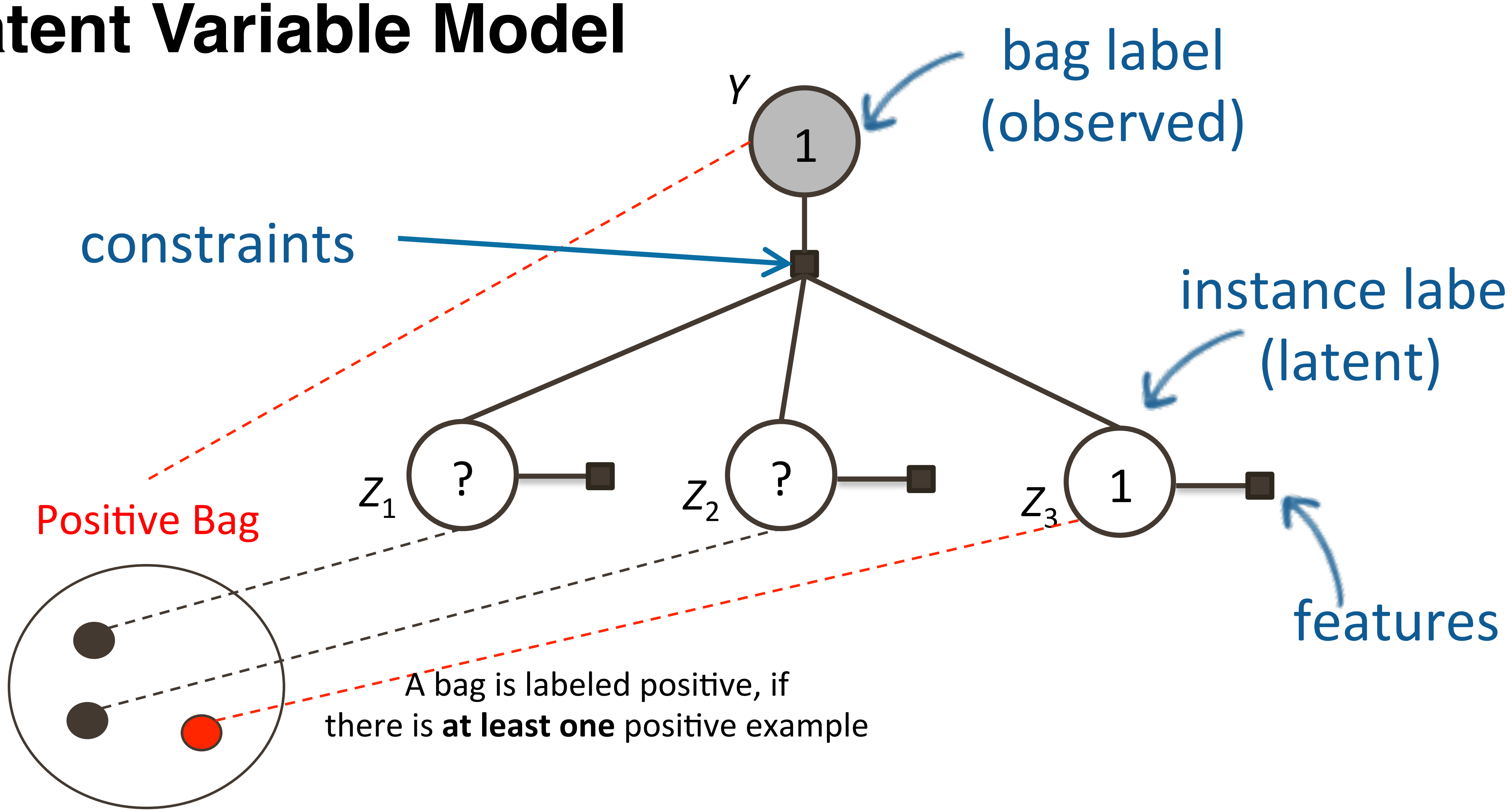
A bag is labeled negative, if
all the examples in it are negative

A bag is labeled positive, if
there is **at least one** positive example

(Dietterich et al., 1997)

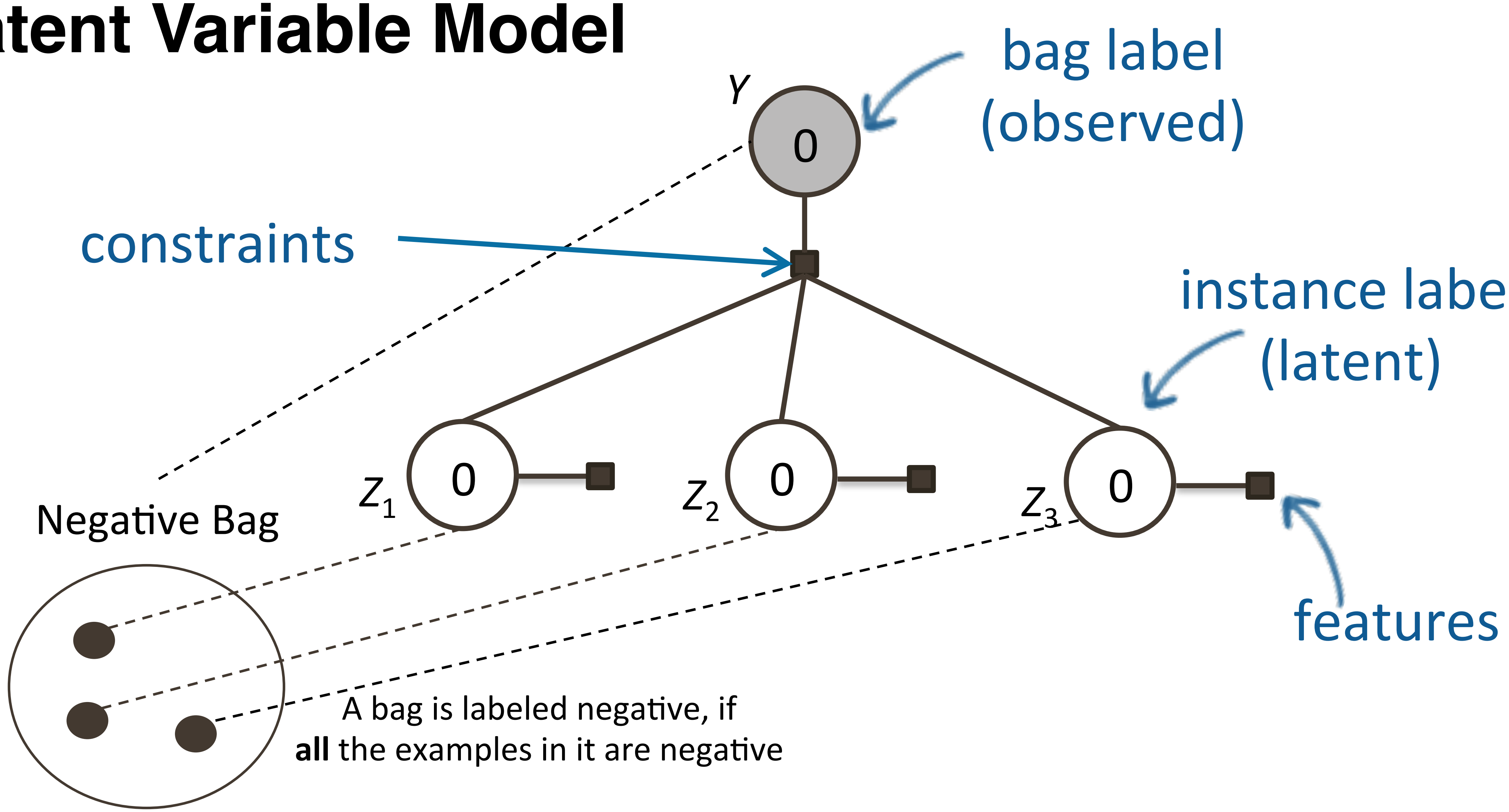
[Mini Tutorial]
Multi-instance Learning

Latent Variable Model



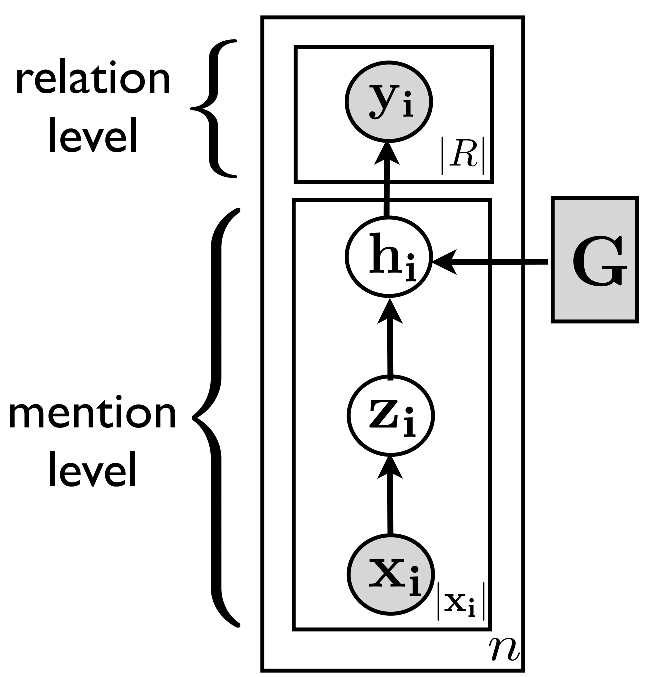
[Mini Tutorial] Multi-instance Learning

Latent Variable Model



[Mini Tutorial]
Multi-instance Learning

Distantly Supervised Information Extraction



1. incomplete knowledge base problem

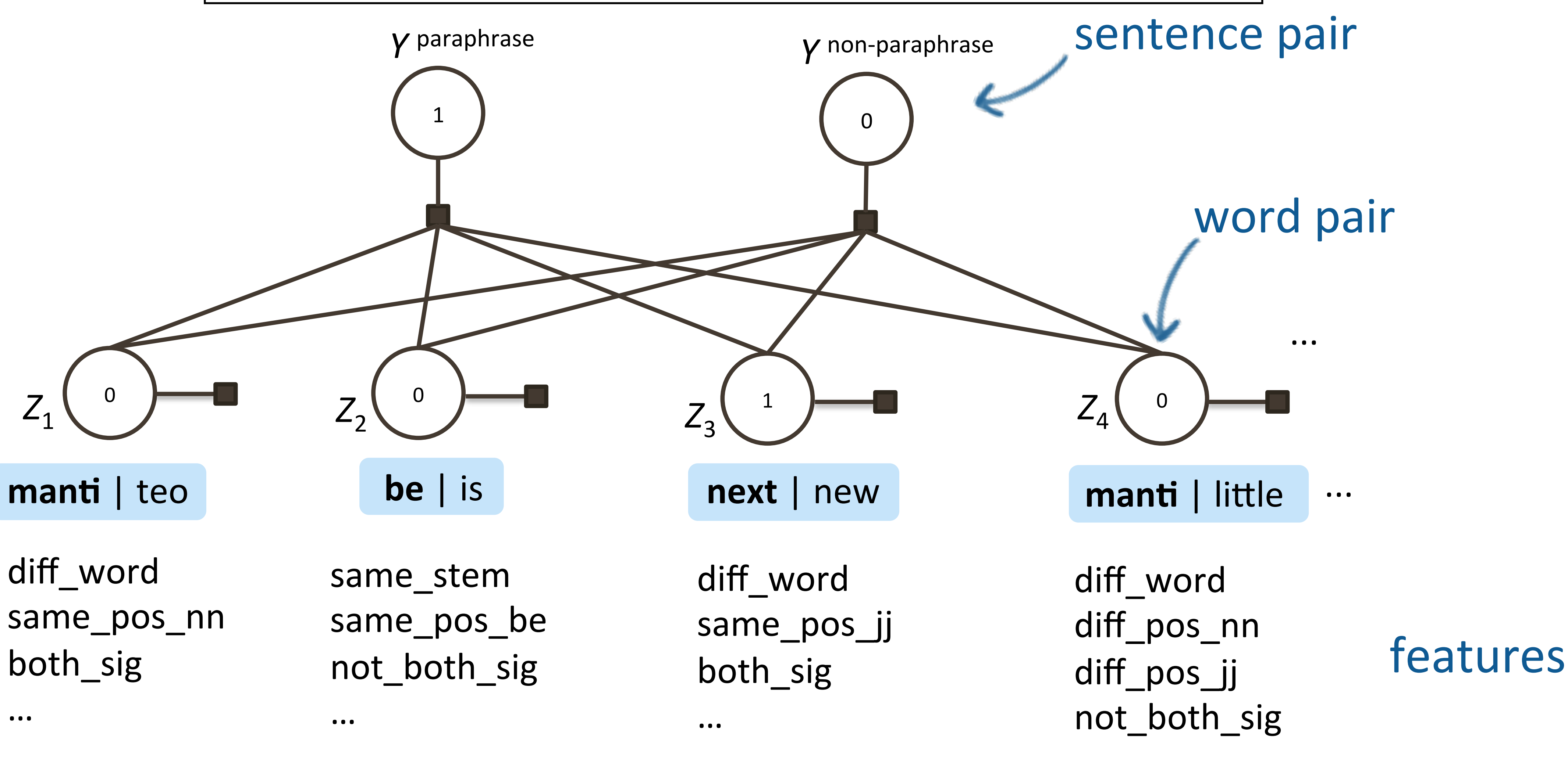
2. distant supervision + human-labeled data

3. IE + IR

Wei Xu, Ralph Grishman, Le Zhao. "Passage Retrieval for Information Extraction using Distant Supervision" In IJCNLP (2011)
Wei Xu, Raphael Hoffmann, Le Zhao, Ralph Grishman. "Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction" In ACL (2013)
Maria Pershina, Bonan Min, Wei Xu, Ralph Grishman. "Infusion of Labeled Data into Distant Supervision for Relation Extraction" In ACL (2014)

[Recap] Multi-instance Learning Paraphrase Model

*Manti bout to be the **next** Junior Seau*
*Teo is the little **new** Junior Seau*



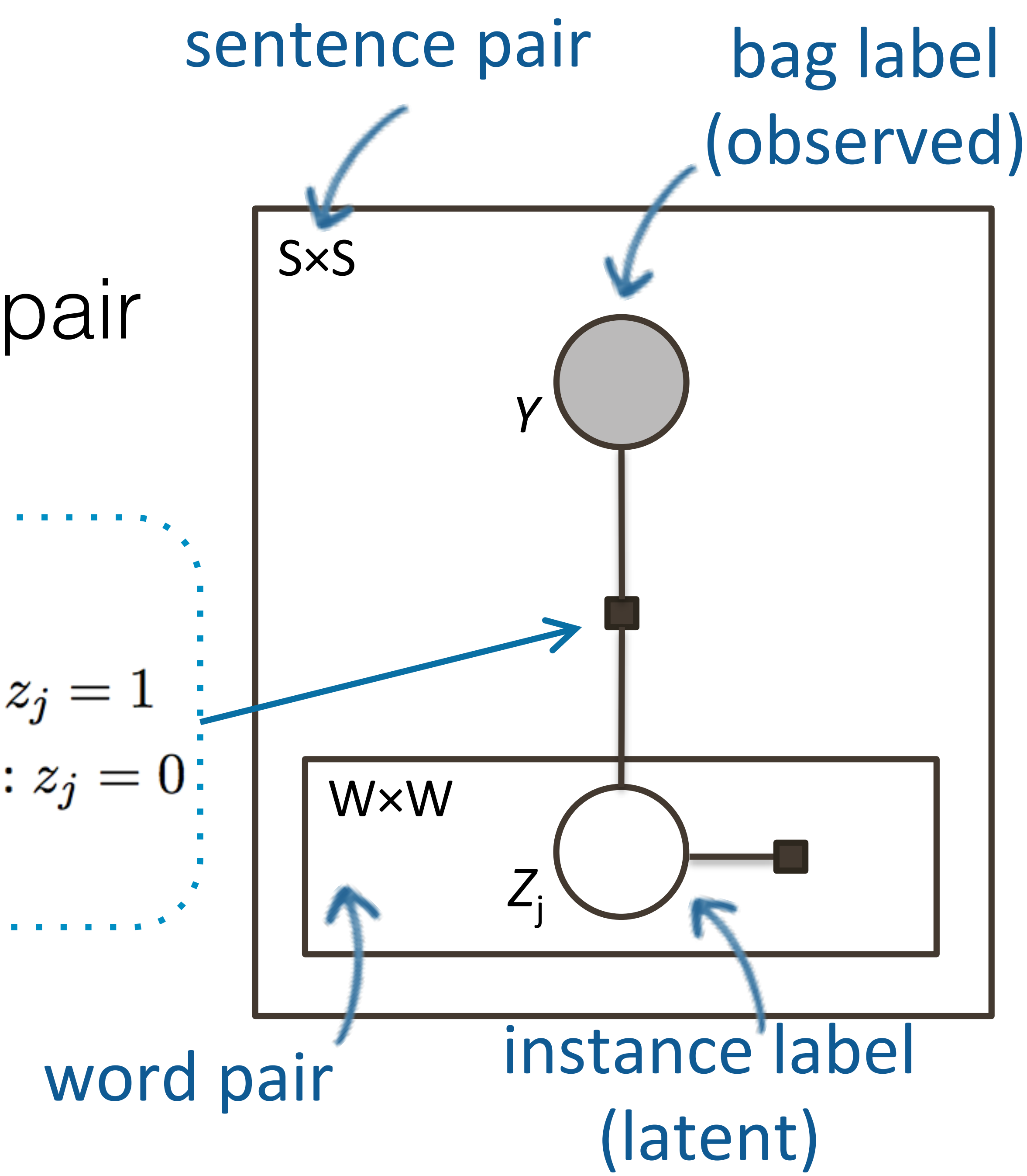
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Joint Word-Sentence Model

Model the assumption:
sentence-level paraphrase
is anchored by at-least-one word pair

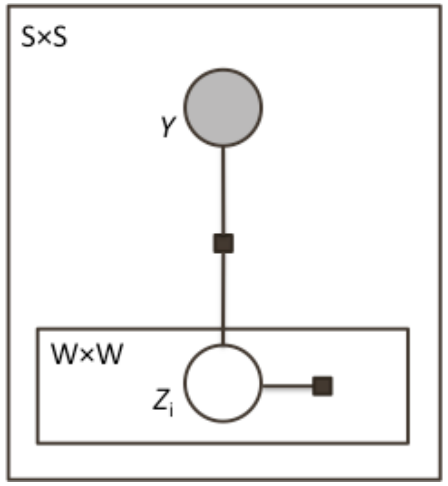
deterministic OR

$$\sigma(\mathbf{z}_i, y_i) = \begin{cases} 1 & \text{if } y_i = true \wedge \exists j : z_j = 1 \\ 1 & \text{if } y_i = false \wedge \forall j : z_j = 0 \\ 0 & \text{otherwise} \end{cases}$$



Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji.
"Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Joint Word-Sentence Model



*i*th sentence pair's label (observed or to be predicated)

*j*th word pair

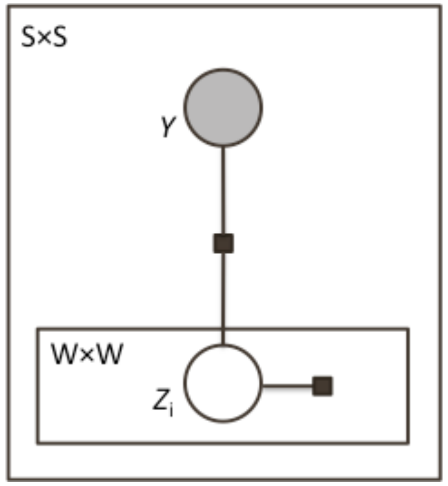
$$P(\mathbf{z}_i, y_i | \mathbf{w}_i; \theta) = \prod_{j=1}^{m} \exp(\theta \cdot f(z_j, w_j)) \times \sigma(\mathbf{z}_i, y_i)$$

parameters features deterministic OR

latent labels for all word pairs in the *i*th sentence pair

Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Learning Algorithm



Objective:
learn the parameters that maximize likelihood over the training corpus

$$\theta^* = \arg\max_\theta P(\mathbf{y}|\mathbf{w};\theta) = \arg\max_\theta \prod_i \sum_{\mathbf{z}_i} \boxed{P(\mathbf{z}_i, y_i|\mathbf{w}_i;\theta)}$$

***i*th training sentence pair**

all possible values of the latent variables

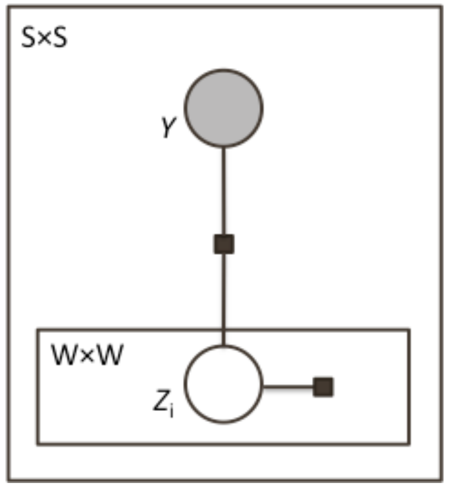
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Learning Algorithm

Perceptron-style Update:

Viterbi approximation + online learning

O(# word pairs)



$$\frac{\partial \log P(\mathbf{y}|\mathbf{w};\theta)}{\partial \theta} \approx \underbrace{\sum_i f(\mathbf{z}_i^*, \mathbf{w}_i)}_{\text{reward correct (conditioned on labels)}} - \underbrace{\sum_i f(\mathbf{z}'_i, \mathbf{w}_i)}_{\text{penalize wrong (ignoring labels)}}$$

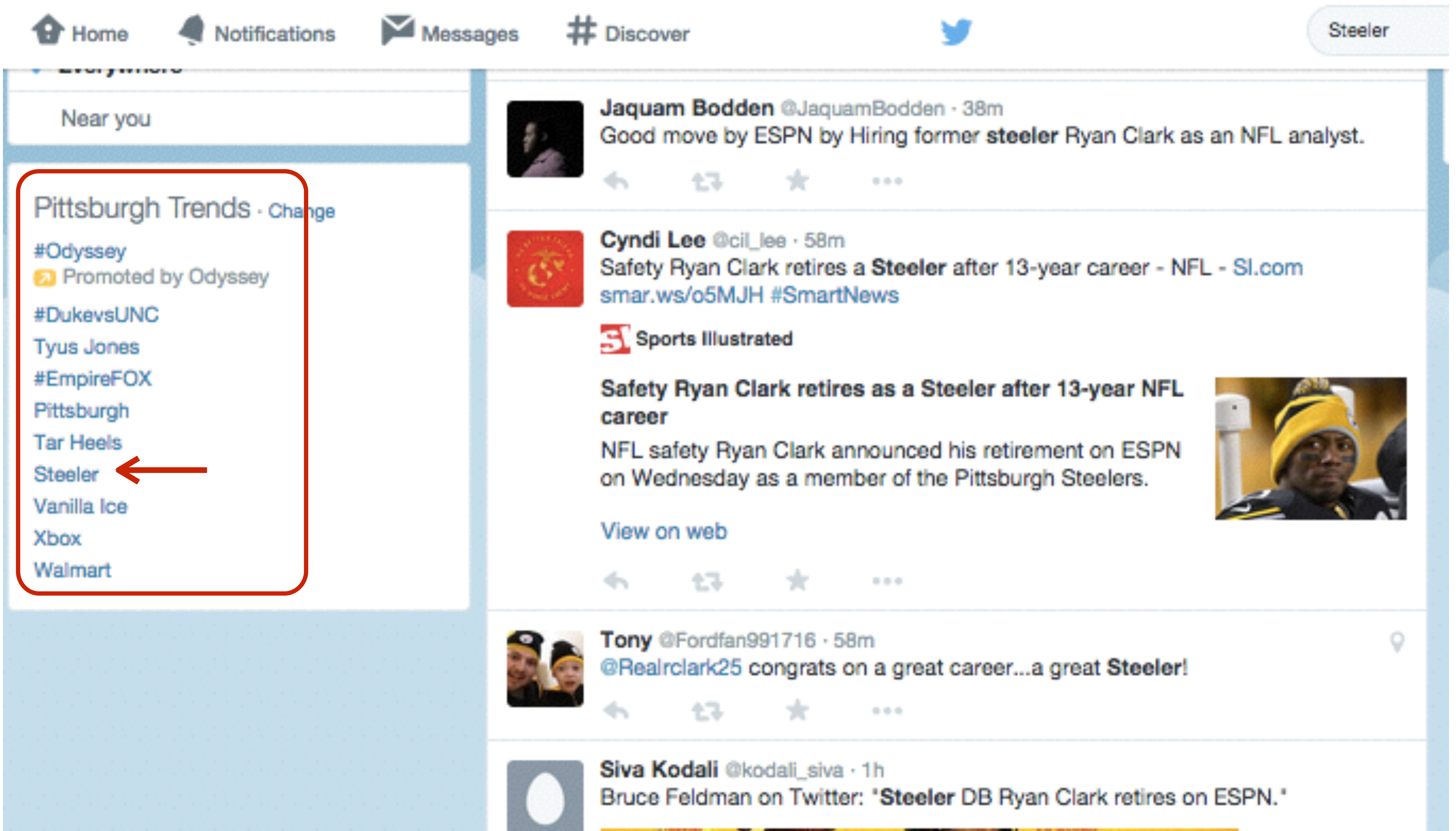
$$\mathbf{z}^* = \arg\max_{\mathbf{z}} P(\mathbf{z}|\mathbf{w}, \mathbf{y}; \theta)$$

$$\mathbf{y}', \mathbf{z}' = \arg\max_{\mathbf{y},\mathbf{z}} P(\mathbf{z}, \mathbf{y}|\mathbf{w}; \theta)$$

Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

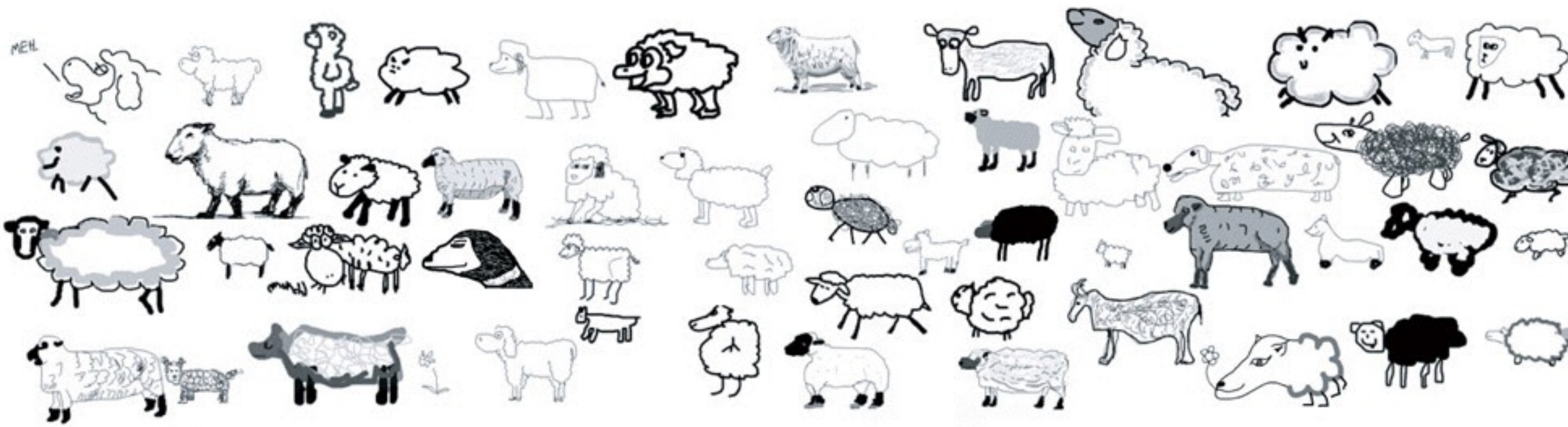
Training Data

Twitter Trends



Annotation

Crowdsourcing



(Courtesy: The Sheep Market by Aaron Koblin)

Annotation

Crowdsourcing

Here Is The Question To You:

Original Sentence: ***Borussia Dortmund advanced to the final***

Select ALL sentences that have similar meaning from below:

- [ ] Borussia Dortmund has clinched their Champions League final spot
- [ ] Real Madrid efforts are not enough as Cinderella Borussia Dortmund advances to the Champions League Final
- [ ] But it s Borussia Dortmund whose heading to Wembley Park
- [ ] Congratulations Borussia Dortmund s going to Wembley

amazon mechanical turk™
Artificial Artificial Intelligence

Wei Xu. “Data-driven Approaches for Paraphrasing Across Language Variations” PhD Thesis, New York University. (2014)

A Problem

only **8%** sentence pairs about the same topic have similar meaning

hurts both quantity and quality

non-experts lower their bars

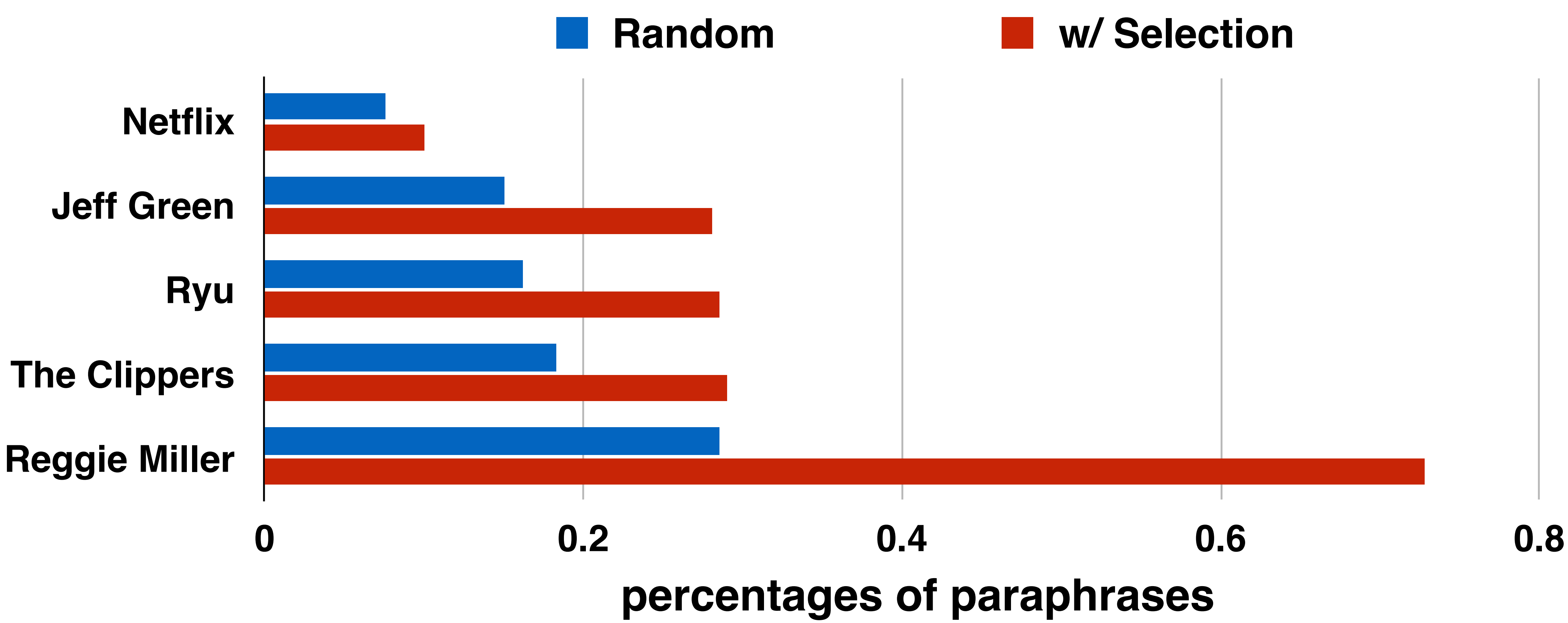
Wei Xu. “Data-driven Approaches for Paraphrasing Across Language Variations” PhD Thesis, New York University. (2014)

Sentence Selection

SumBasic Algorithm

8% → 16%

$$Salience(s) = \sum_{w_i \in s} \frac{P(w_i)}{|w_i | w_i \in s|}$$

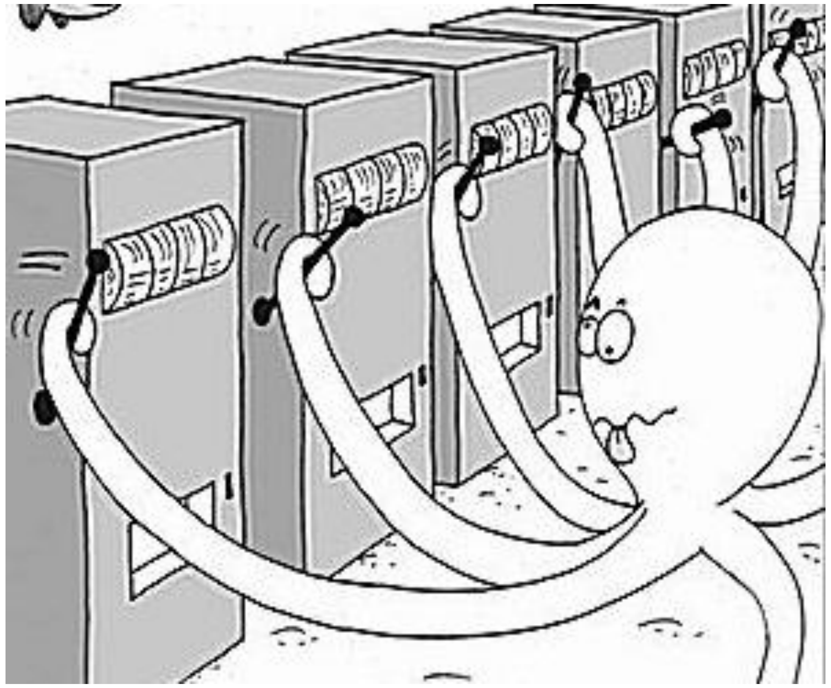


Wei Xu. "Data-driven Approaches for Paraphrasing Across Language Variations" PhD Thesis, New York University. (2014)

Topic Selection

Multi-Armed Bandits

16% → 34%



$$\max \sum_{\{i | r_i(t_0) > 0\}} \hat{\mu}_i(t_0) r_i(t_1)$$

$$\text{s.t.} \quad \sum_i c_i r_i(t_1) \leq (1 - \epsilon)B, \forall i : 0 \leq r_i(t_1) \leq l - r_i(t_0).$$

Wei Xu. “Data-driven Approaches for Paraphrasing Across Language Variations” PhD Thesis, New York University. (2014)

Twitter Paraphrase Dataset

18,762 sentence pairs labeled
cost only $200

important but difficult to obtain

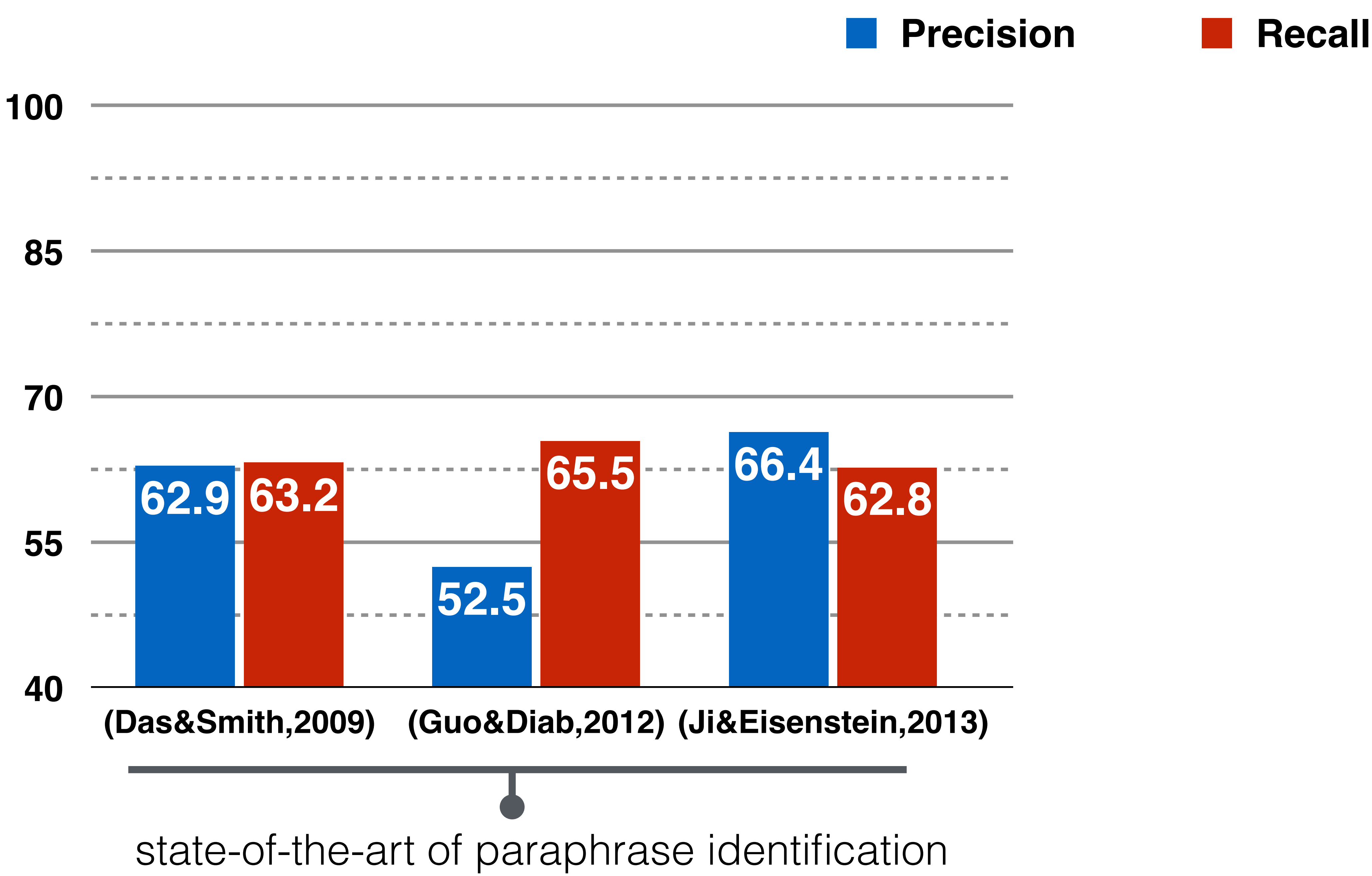
1/3 paraphrase, 2/3 non-paraphrase (very balanced)

including a very broad range of paraphrases:
synonyms, misspellings, slang, acronyms and colloquialisms

Wei Xu. “Data-driven Approaches for Paraphrasing Across Language Variations” PhD Thesis, New York University. (2014)

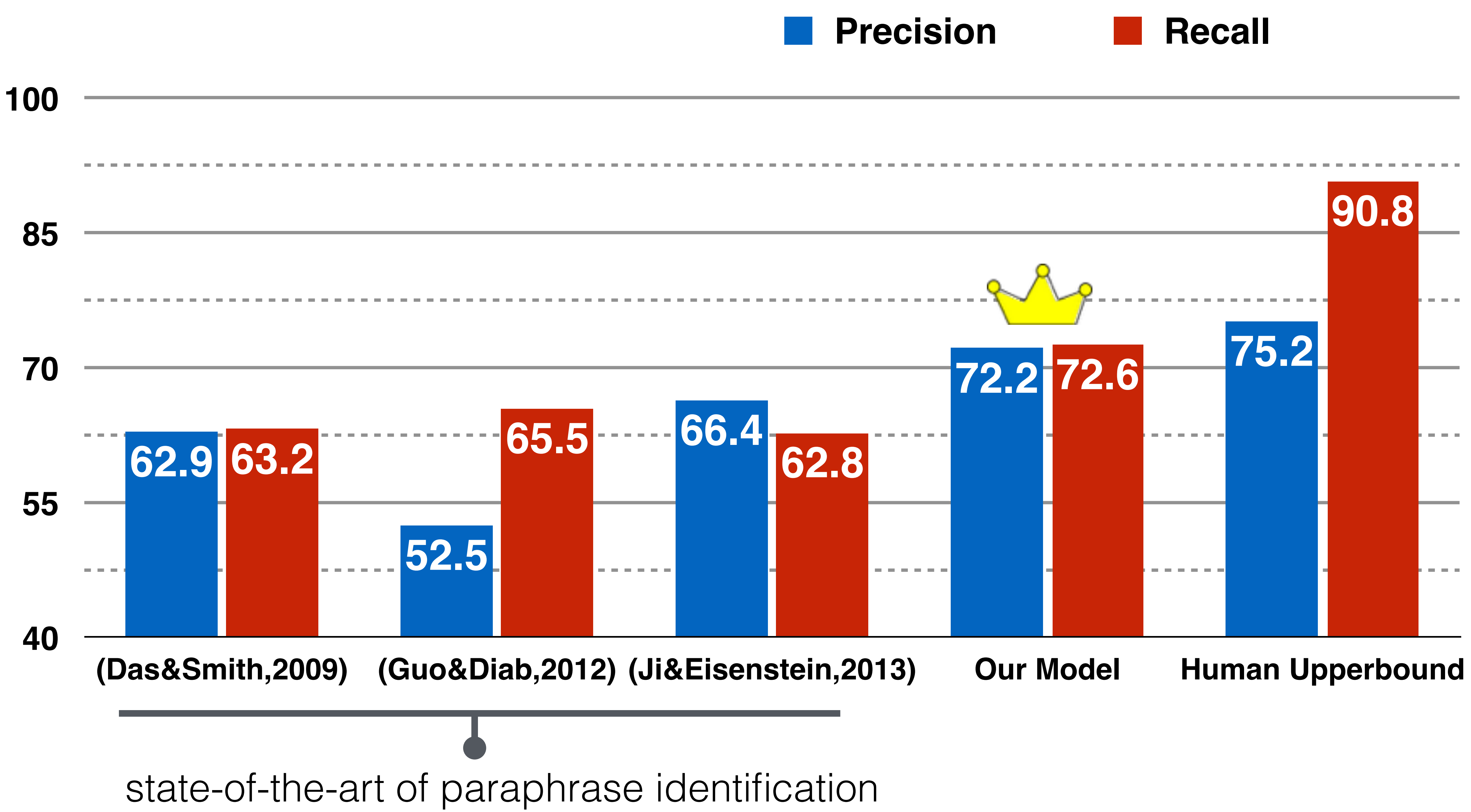
Performance

Performance



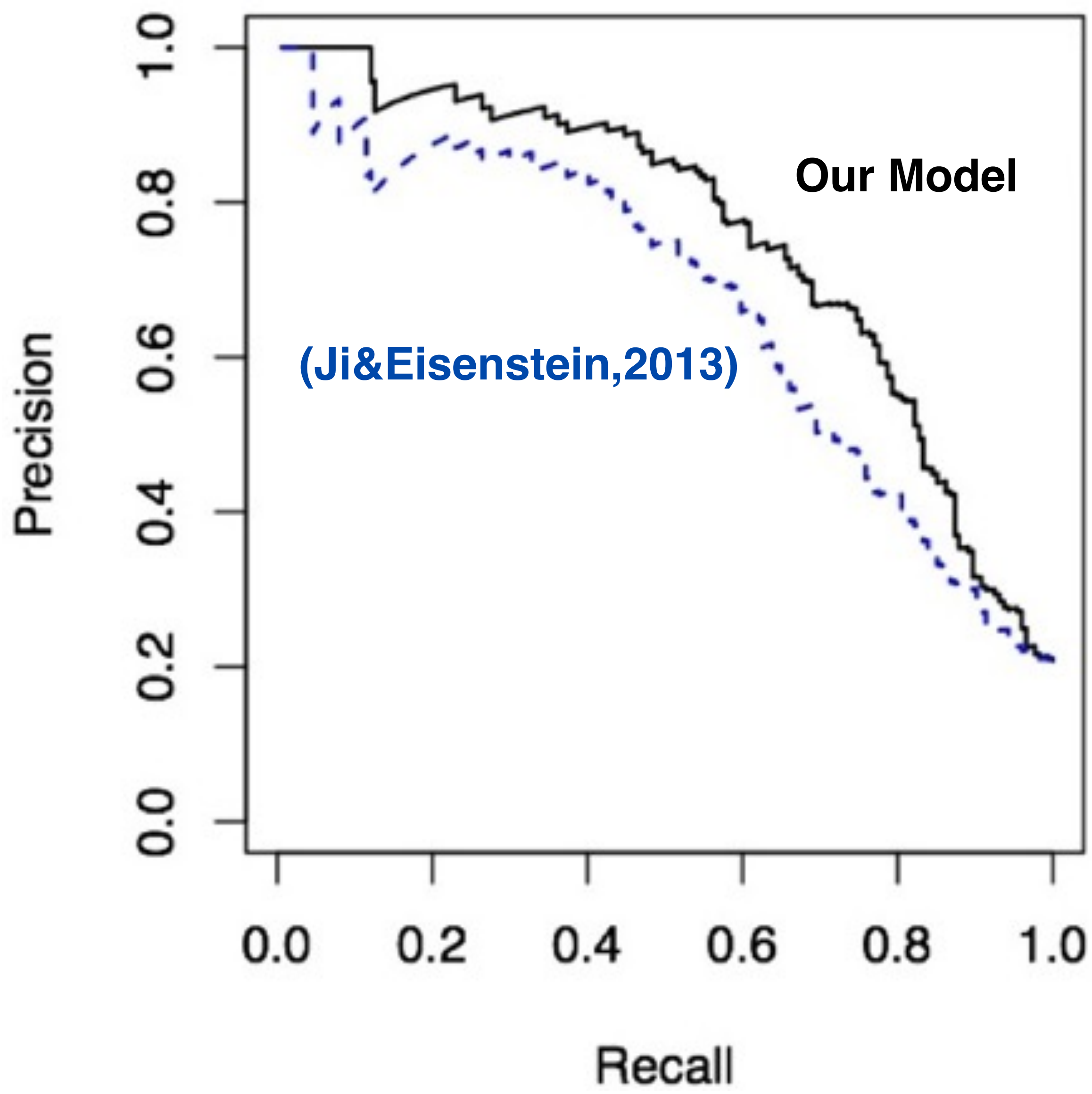
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

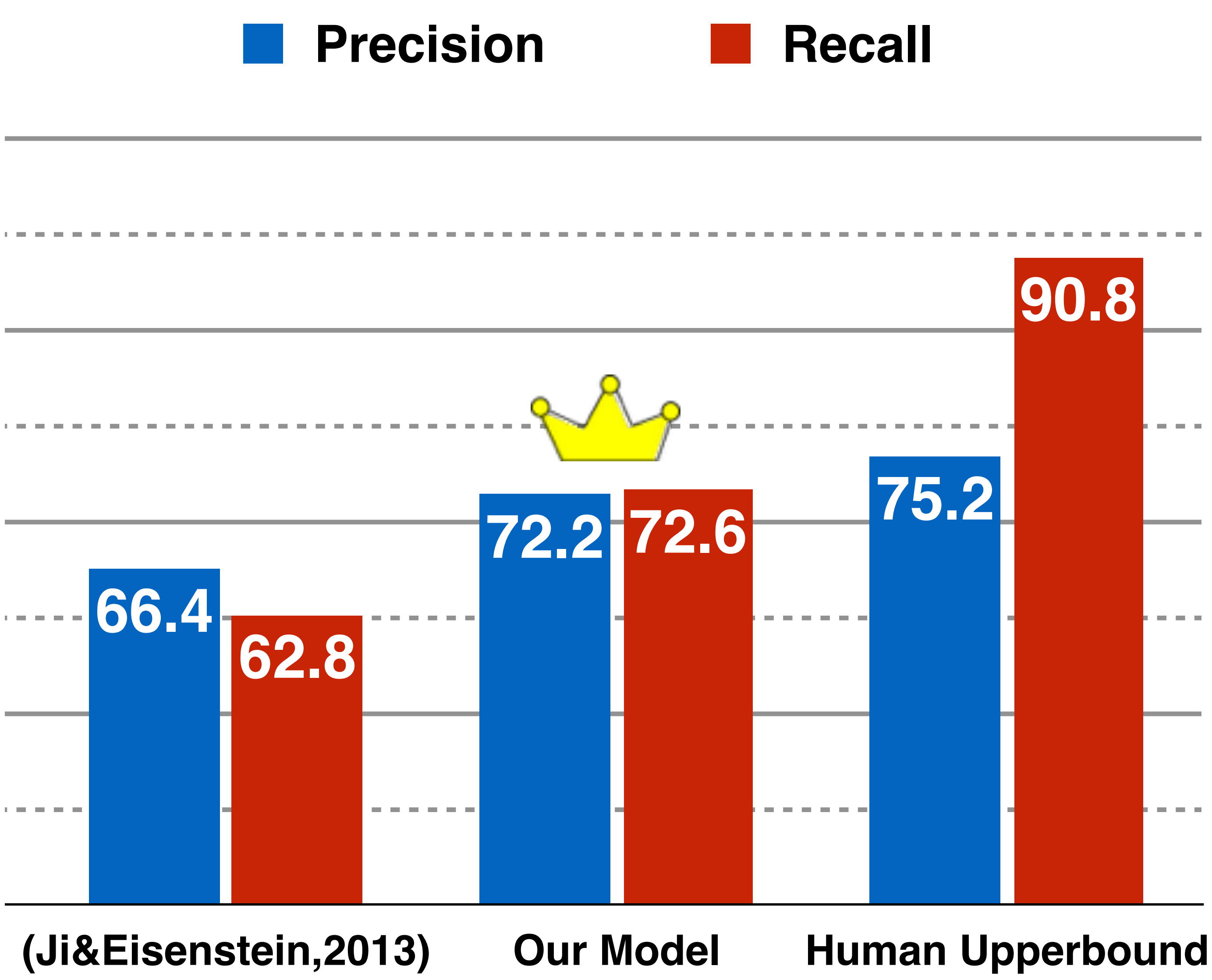
Performance



Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji.
"Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Performance





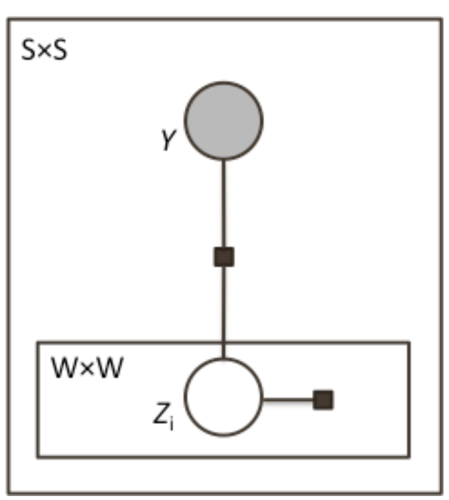
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Innovations

That boy Brook Lopez with a deep **3**
brook lopez hit a **3**

Yes!

Multi-instance Learning Paraphrase Model (MultiP)



- Twitter's big data stream
- potential beyond Twitter and English
- joint sentence-word alignment
- extensible latent variable model

(a lot of space for future work)

Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Impact

SemEval 2015

Shared Task: Paraphrase and Semantic Similarity in Twitter
19 research groups participated (100+ requested the data)

Wei Xu, Chris Callison-Burch, Bill Dolan.
"SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter" In SemEval (2015)

thank you very much

thank u 4 ur time

Thank you

thanks

thanking you

gratitude

appreciate it

thx

Follow @cocoweixu

3x

say thanks

Instructor: Wei Xu

www.cis.upenn.edu/~xwe/

tyvm

thnx

Course Website: socialmedia-class.org

wawwww thankkkkkkkkkkk you alottttttttttttt!

thanks a lot

am grateful