Learning and Generating Paraphrases

From Twitter and Beyond

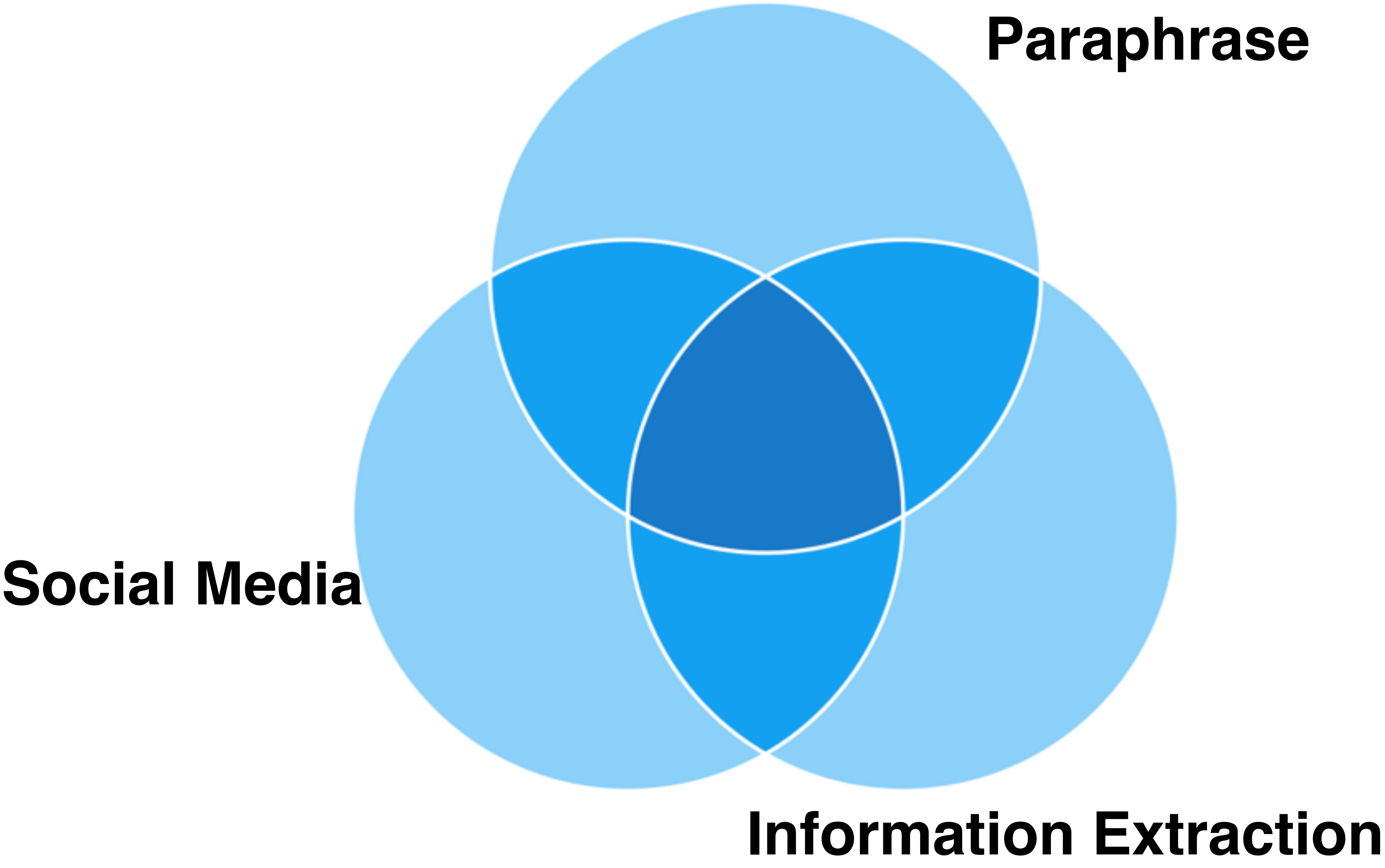
Wei Xu

Computer and Information Science
University of Pennsylvania

Follow @cocoweixu

Guest Lecture @ Penn MT class April-2-2015

Research Overview



Paraphrase

Social Media

Information Extraction

TACL 15

NAACL 15

TACL 14

ACL 14

ACL 13

BUCC 13

LSAM 13

COLING 12

IJCNLP 11

EMNLP 11

ACL 06

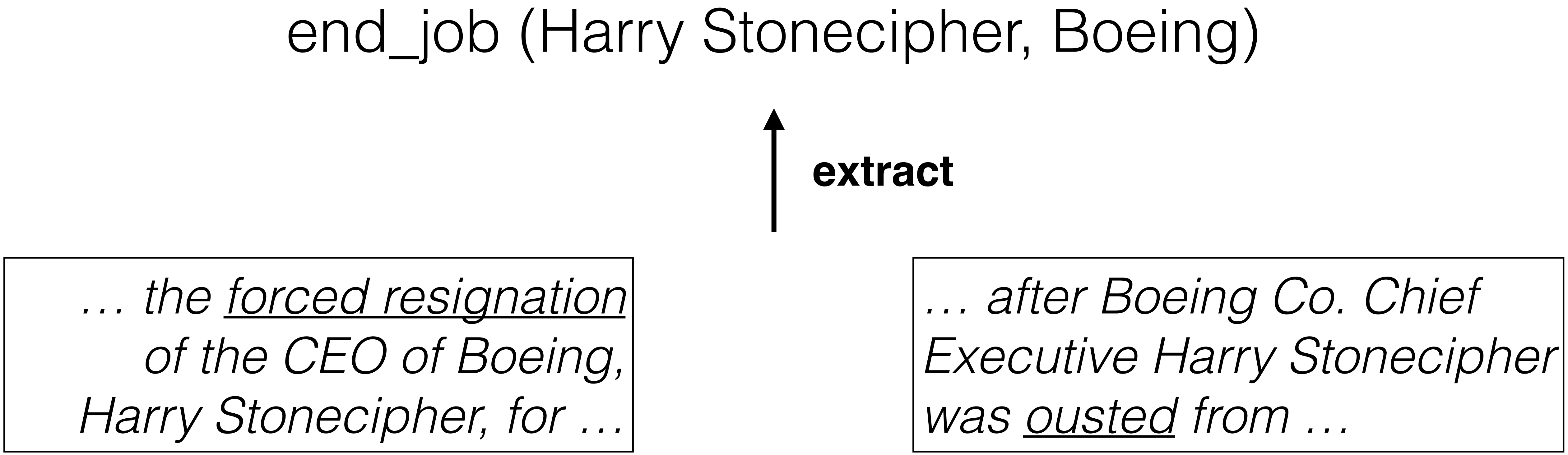
Paraphrase

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

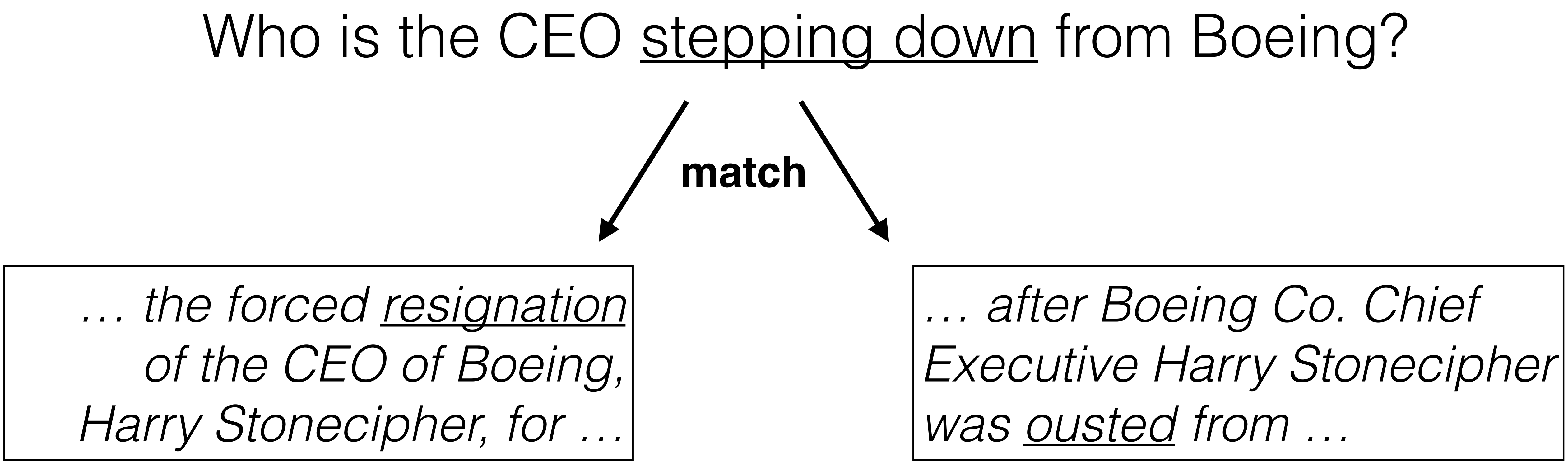
… the forced resignation of the CEO of Boeing, Harry Stonecipher, for …

… 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?

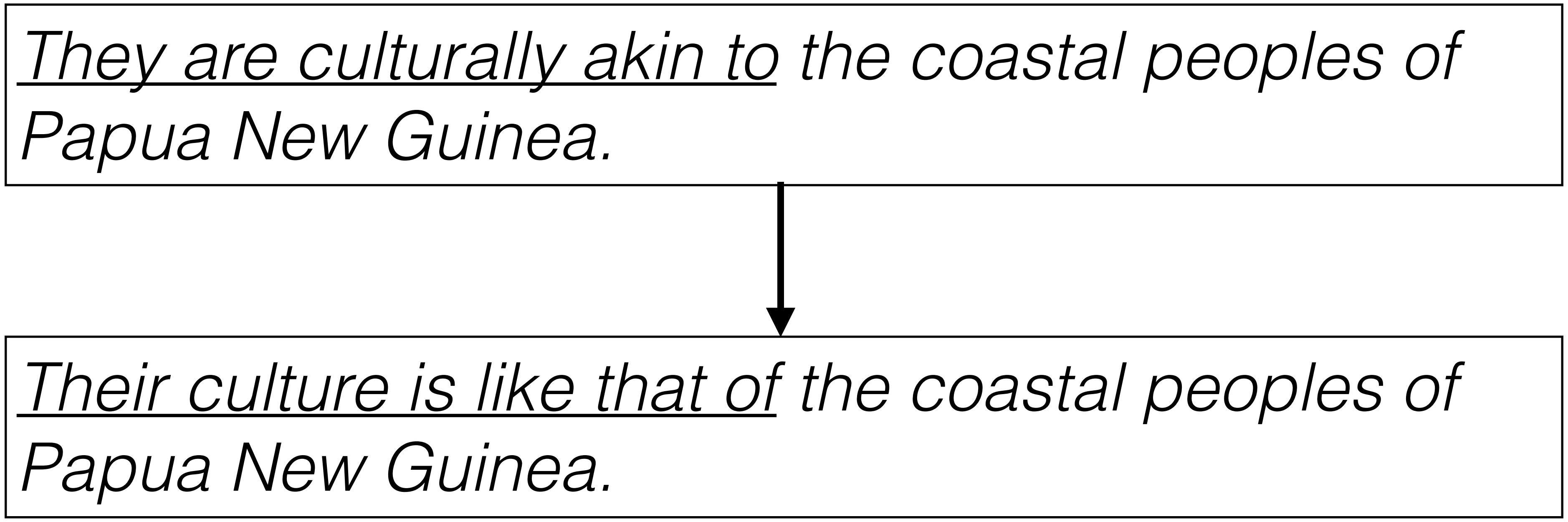
match

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

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

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” to appear in 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)

Previous Work

Numerous publications on paraphrase identification, extraction, generation and various applications

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

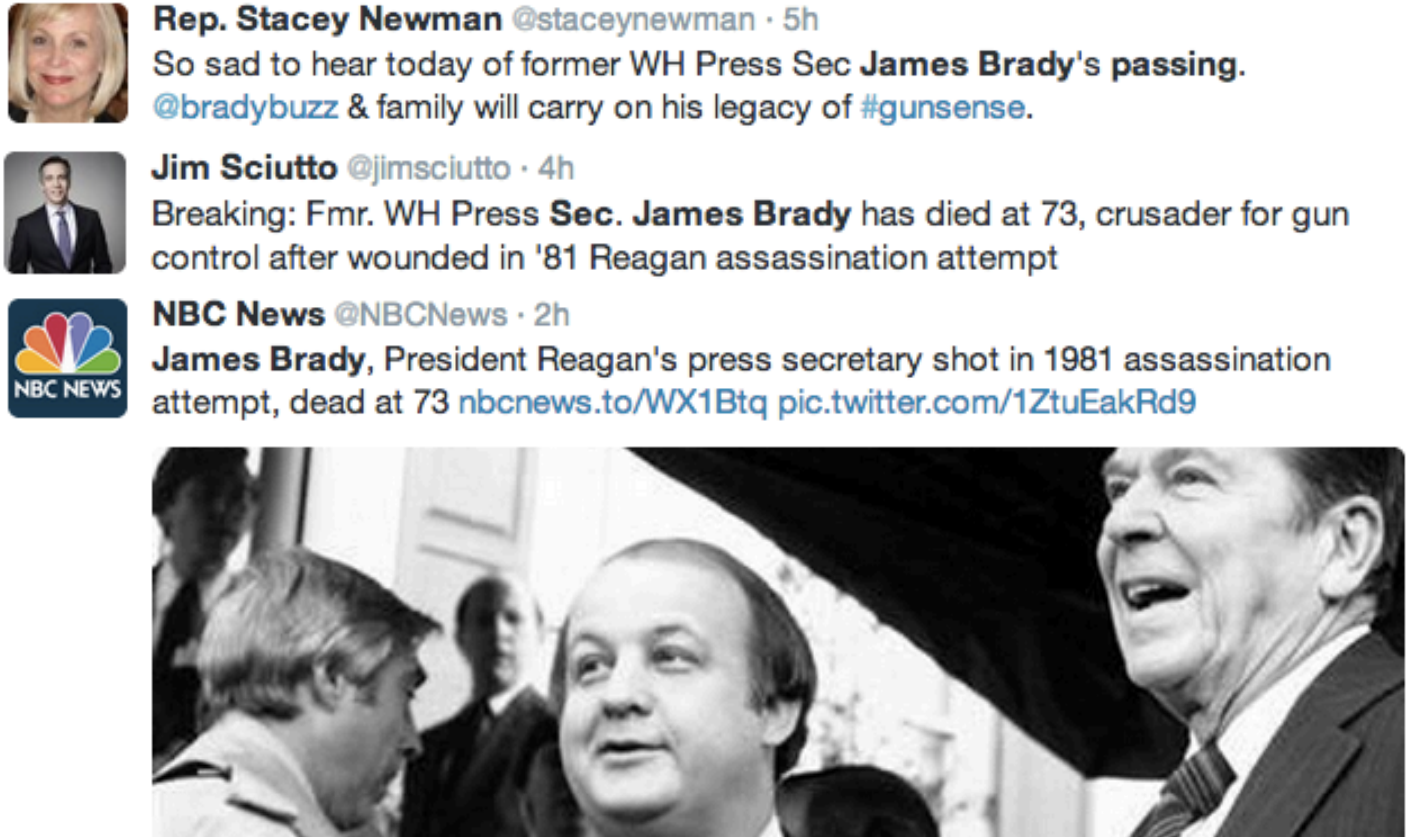
Previous Work

Daily Mail
HIJACKING OF A VERY MIDDLE CLASS PROTEST
theguardian
'This is just the beginning'
THE TIMES
'Thuggish and disgraceful'
THE INDEPENDENT
FINANCIAL TIMES
THE G20 SUMMIT
Ireland's borrowing costs rise sharply
The Daily Telegraph
Work-shy to lose benefits for 3 years
Embarrassed police chiefs caught out by

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)

Twitter as a new resource



Rep. Stacey Newman @staceynewman · 5h

So sad to hear today of former WH Press Sec **James Brady**'s **passing**. @bradybuzz & family will carry on his legacy of #gunsense.

Jim Sciutto @jimsciutto · 4h

Breaking: Fmr. WH Press **Sec**. **James Brady** has died at 73, crusader for gun control after wounded in '81 Reagan assassination attempt

NBC News @NBCNews · 2h

James Brady, President Reagan's press secretary shot in 1981 assassination attempt, dead at 73 nbcnews.to/WX1Btq pic.twitter.com/1ZtuEakRd9

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

Twitter as a powerful resource

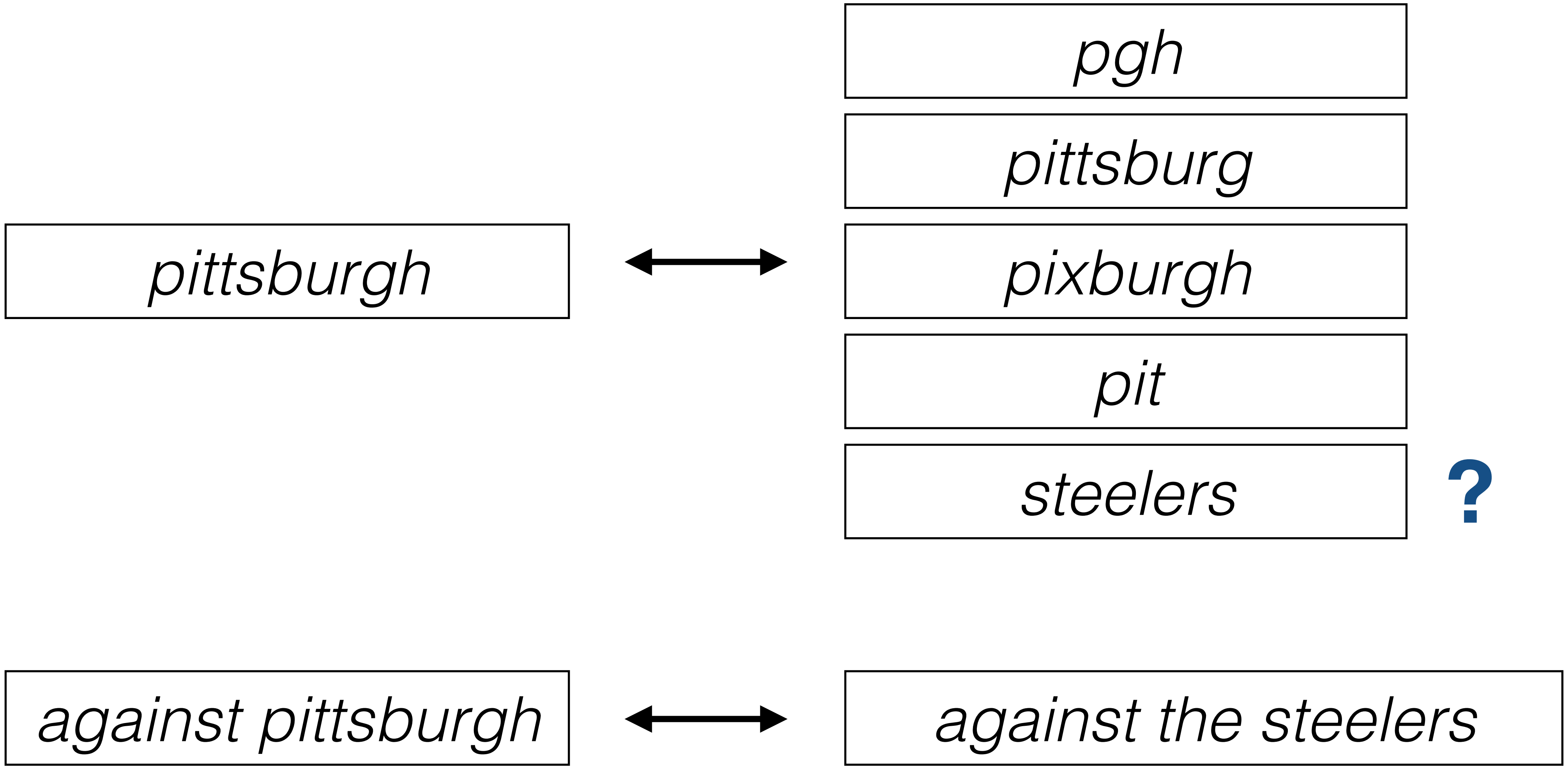
thousands of users
talk about both big and micro events
using formal, informal, erroneous language

Very diverse!

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

Enables new applications

Information Retrieval



pittsburgh ⟷ *pgh*

pittsburg

pixburgh

pit

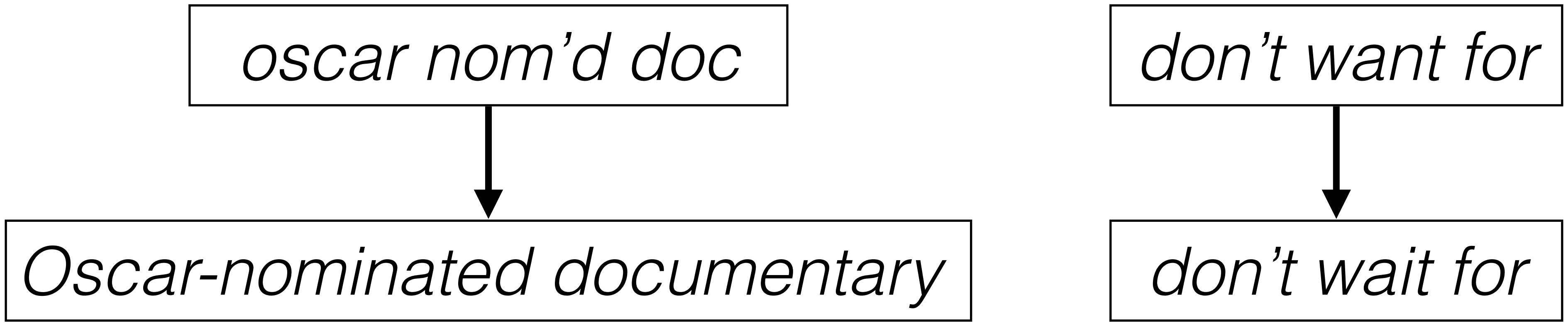
steelers ?

against pittsburgh ⟷ *against the steelers*

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

Enables new applications

Noisy Text Normalization



oscar nom'd doc

Oscar-nominated documentary

don't want for

don't wait for

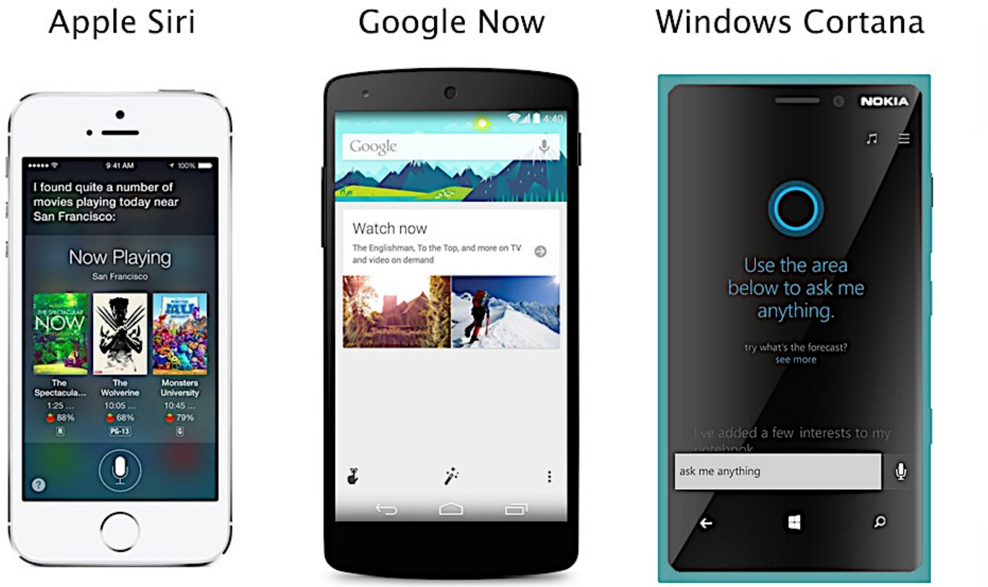
Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao. "Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-gram Models" In EMNLP (2011)

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?*



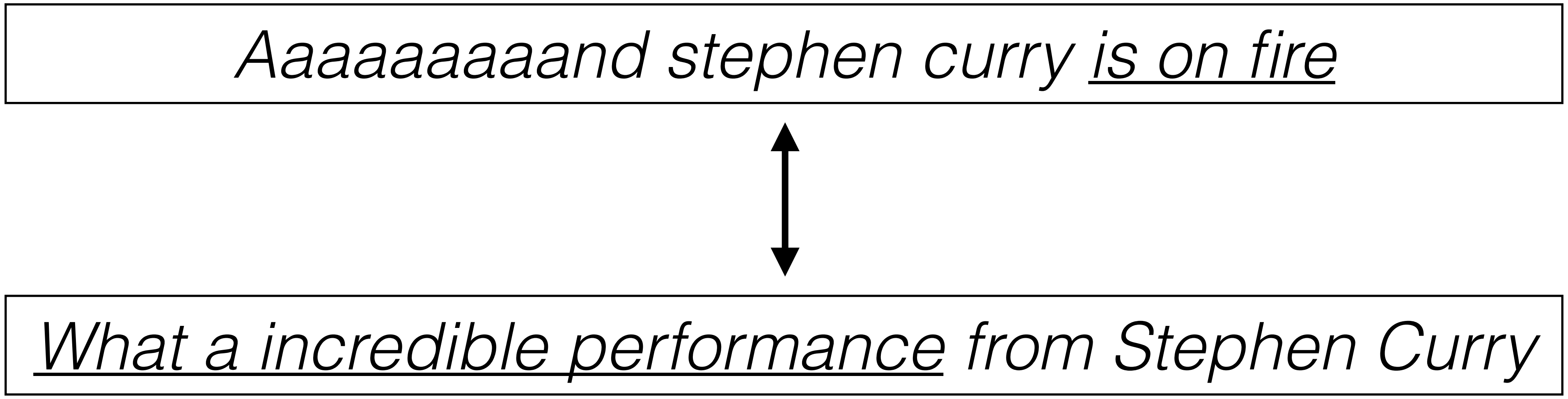
… (21 different ways)

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 is on fire

What a incredible performance from Stephen Curry

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

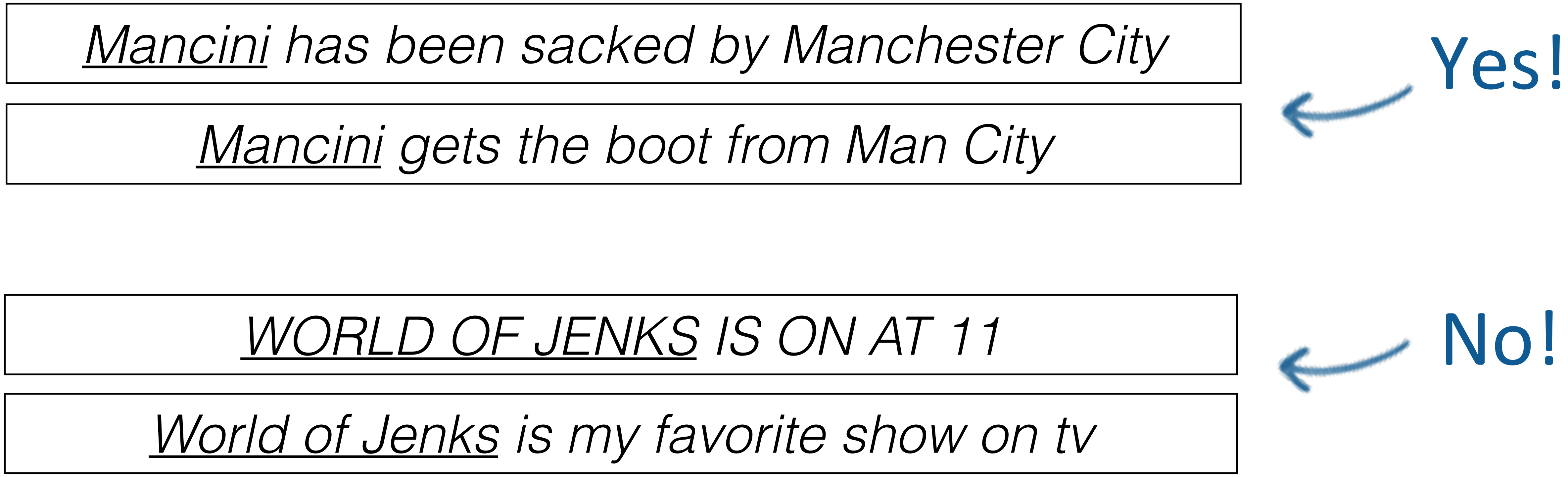
This netsbulls game is too good

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

- 1242 tweet pairs, tracking celebrity & hashtags (Zanzotto, Pennacchiotti and Tsioutsiouliklis, 2011)

- named entity + date (Xu, Ritter and Grishman, 2013)

- bilingual posts (Ling, Dyer, Black and Trancoso, 2013)

Design a Model

Train it on data

A Challenge

Mancini has been sacked by Manchester City

Mancini gets the boot from Man City

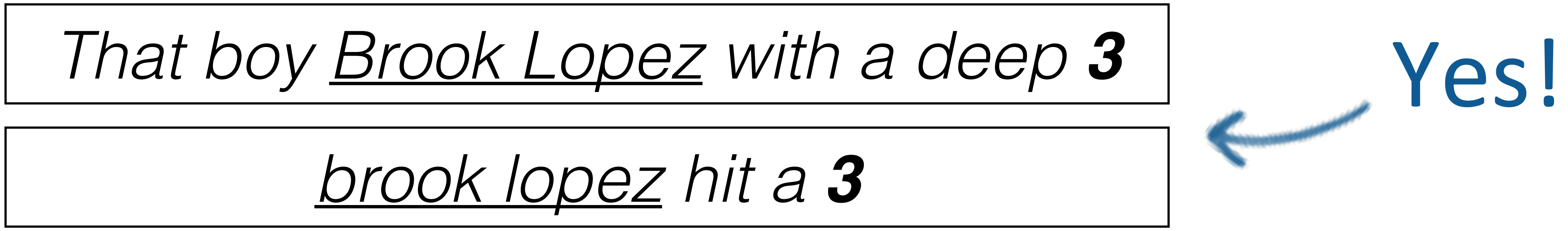
very short
lexically divergent

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

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

Iron Man ***3*** *was brilliant fun*

Iron Man ***3*** *tonight see what this is like*

No!

Solution:

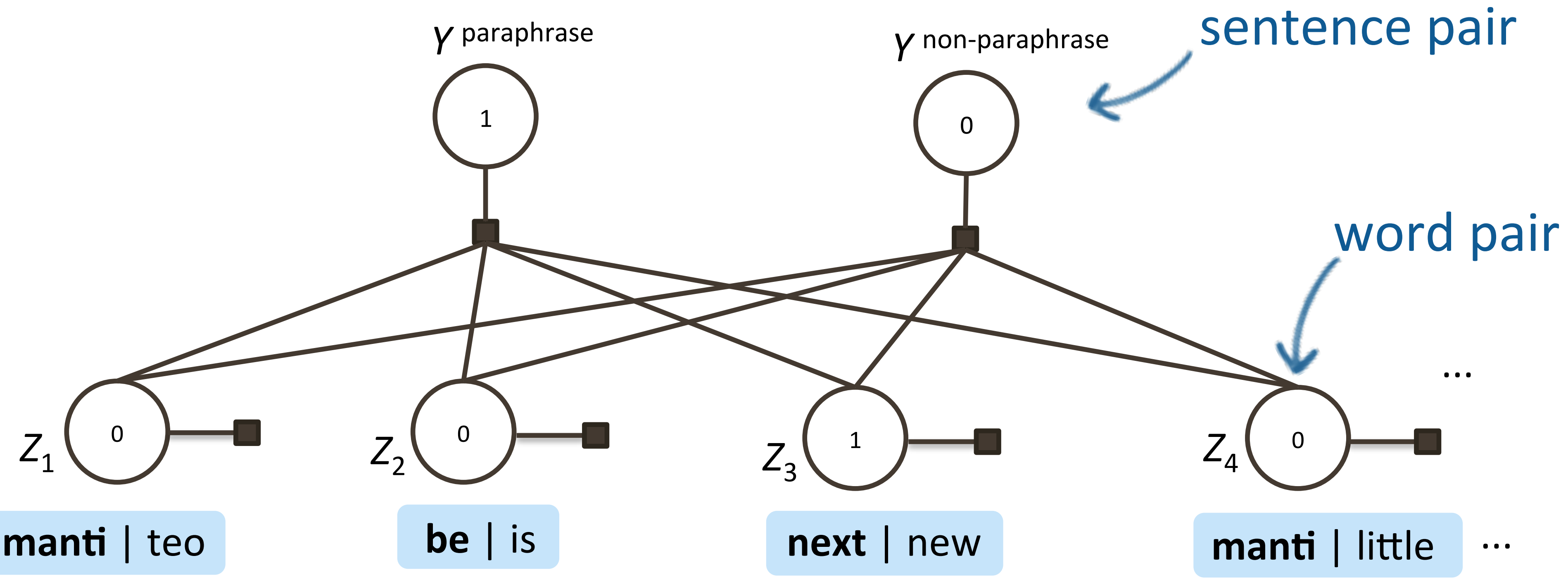
a discriminative model using features at word-level

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

Multi-instance Learning Paraphrase Model

Manti bout to be the ***next*** *Junior Seau*

Teo is the little ***new*** *Junior Seau*



sentence pair

word pair

manti | teo

be | is

next | new

manti | little

...

diff_word
same_pos_nn
both_sig
…

same_stem
same_pos_be
not_both_sig
…

diff_word
same_pos_jj
both_sig
…

diff_word
diff_pos_nn
diff_pos_jj
not_both_sig
…

features

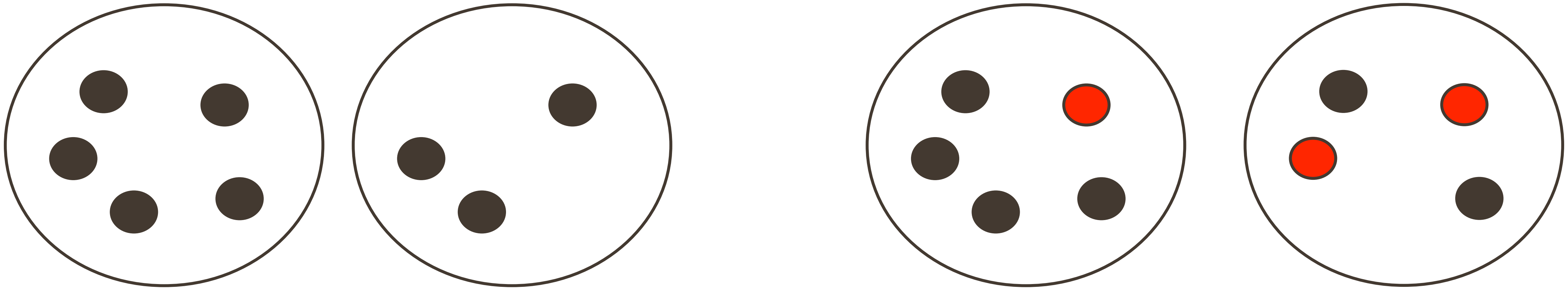
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. “Extracting Lexically Divergent Paraphrases from Twitter” In TACL (2014)

[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



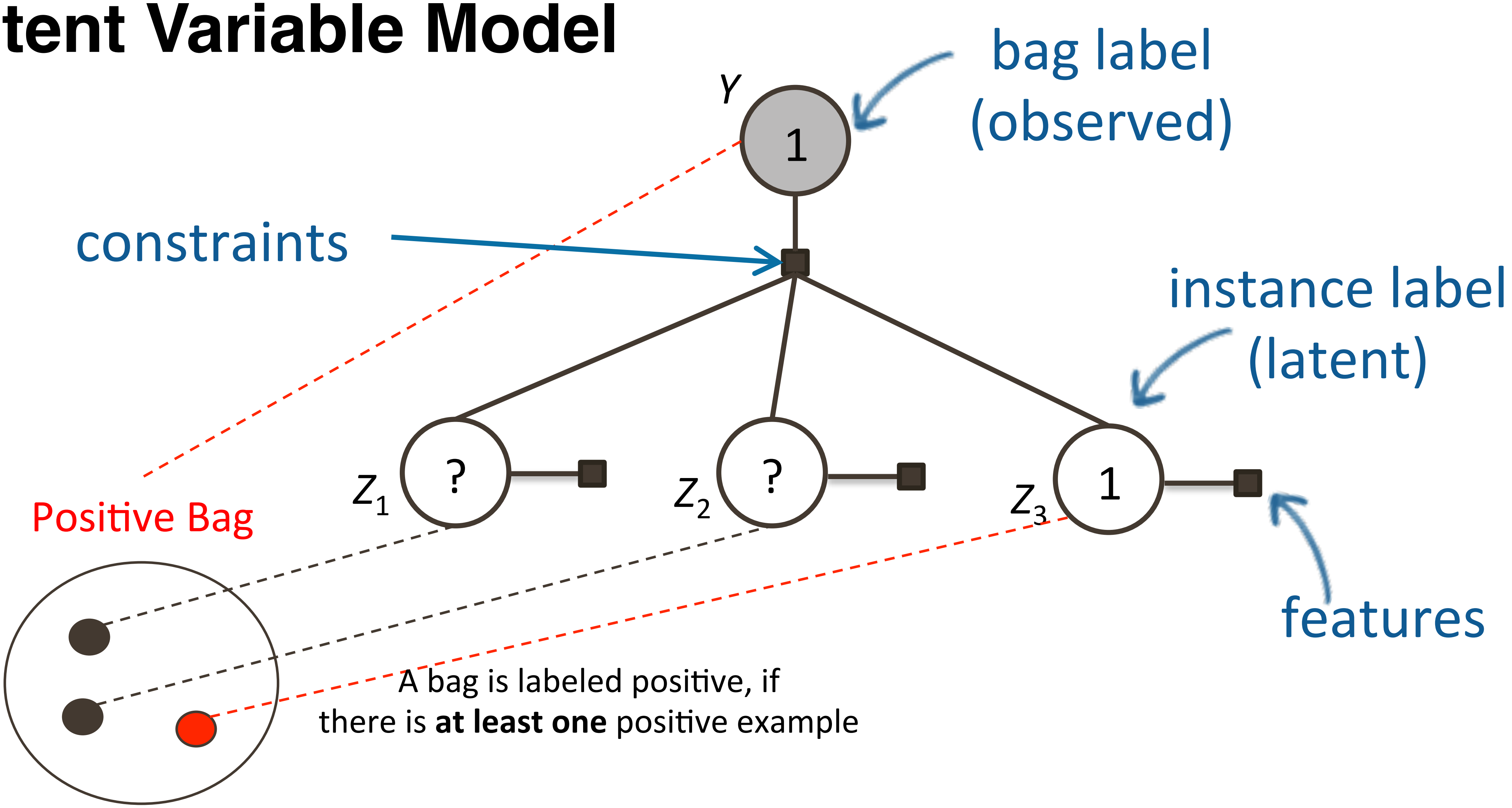
A bag is labeled negative, if
all the examples in it are negative

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

(Dietterich et al., 1997)

[Mini Tutorial]
Multi-instance Learning

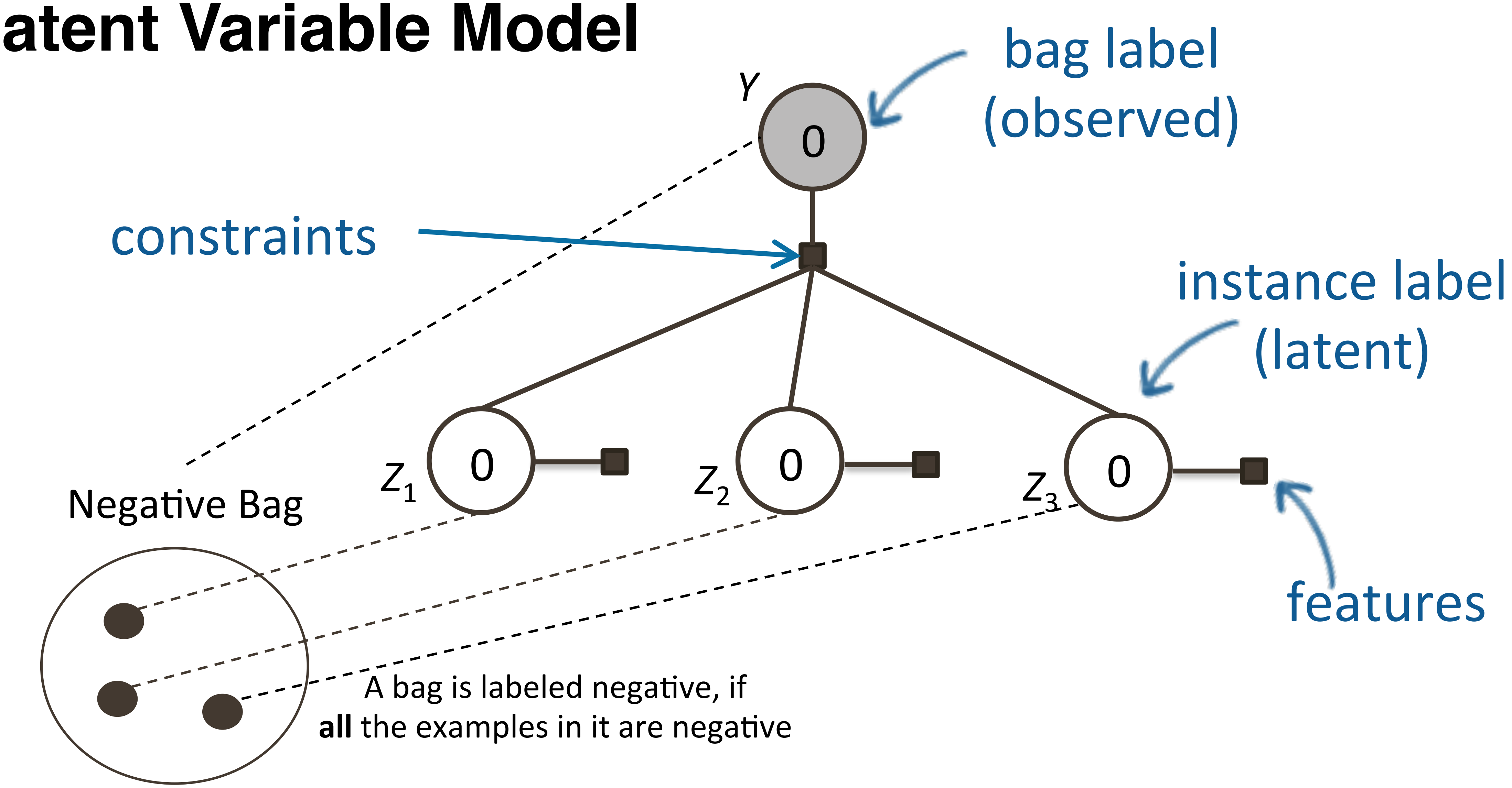
Latent Variable Model



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

[Mini Tutorial]
Multi-instance Learning

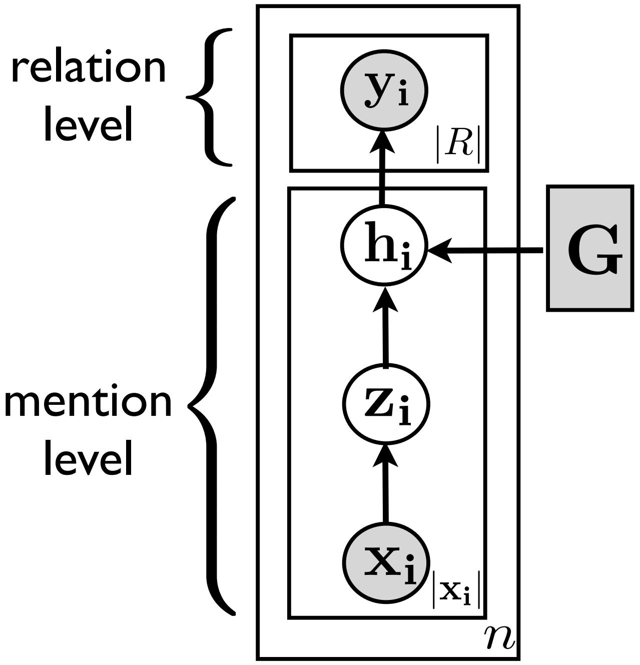
Latent Variable Model



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

[Mini Tutorial]
Multi-instance Learning

Distantly Supervised Information Extraction



1. incomplete knowledge base problem

2. distant supervision + human-labeled data

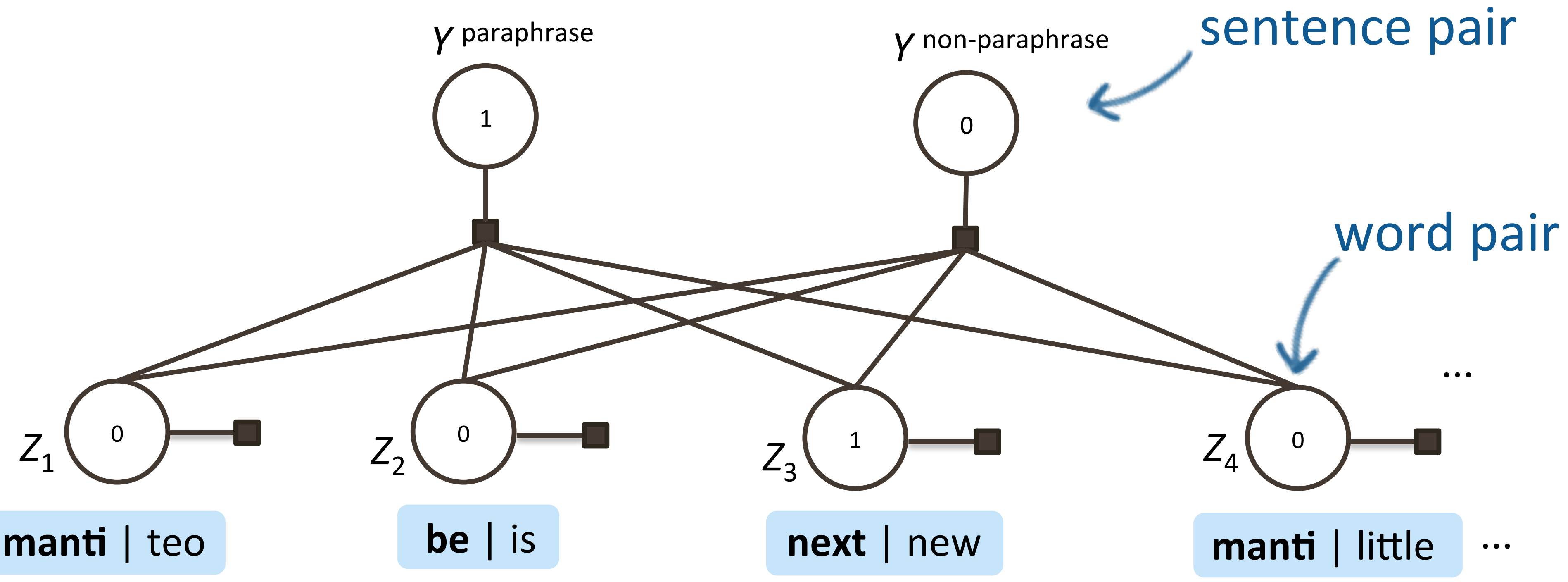
3. IE + IR

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*



diff_word
same_pos_nn
both_sig
...

same_stem
same_pos_be
not_both_sig
...

diff_word
same_pos_jj
both_sig
...

diff_word
diff_pos_nn
diff_pos_jj
not_both_sig
...

features

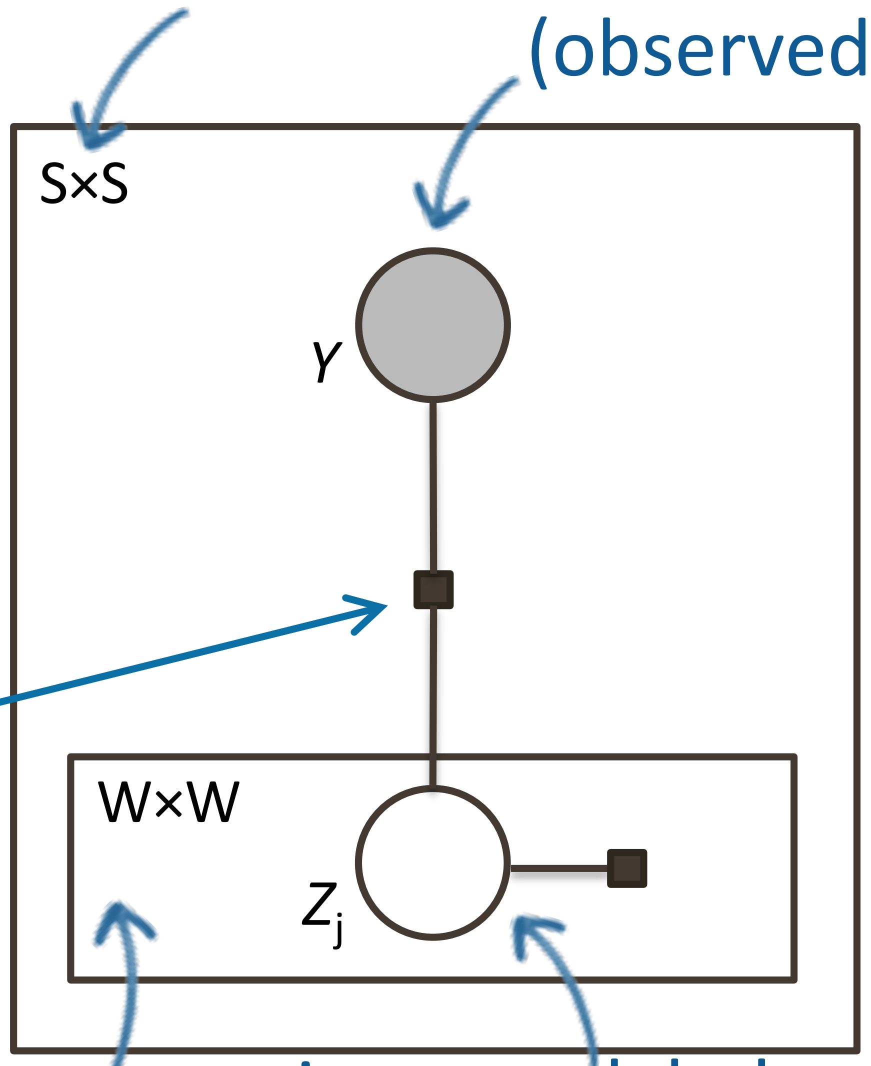
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Joint Word-Sentence Model

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

deterministic OR

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



sentence pair

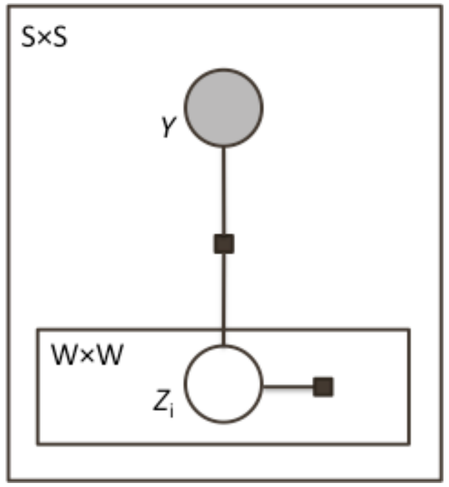
bag label
(observed)

word pair

instance label
(latent)

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

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

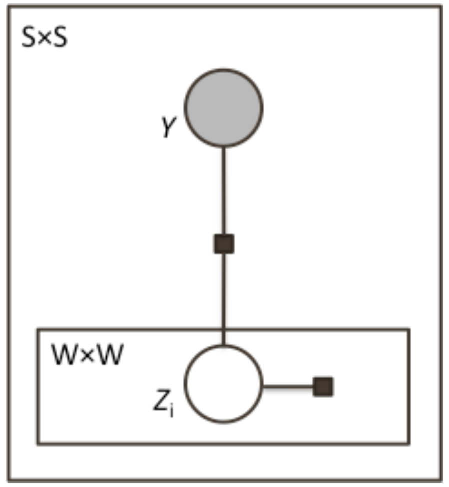
parameters features deterministic OR

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

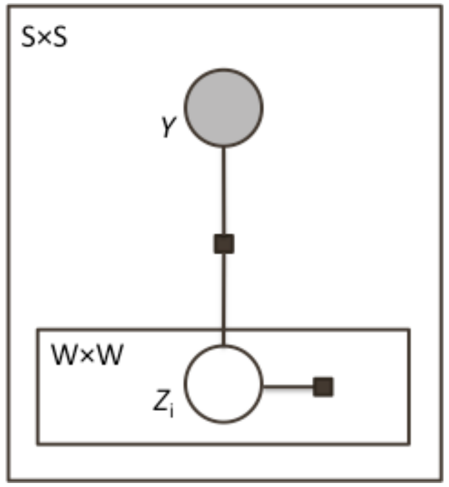
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Learning Algorithm

Perceptron-style Update:

Viterbi approximation + online learning

O(# word pairs)



$$\frac{\partial \log P(\mathbf{y}|\mathbf{w};\theta)}{\partial \theta} \approx \sum_i f(\mathbf{z}_i^*, \mathbf{w}_i) - \sum_i f(\mathbf{z}'_i, \mathbf{w}_i)$$

reward correct (conditioned on labels)

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

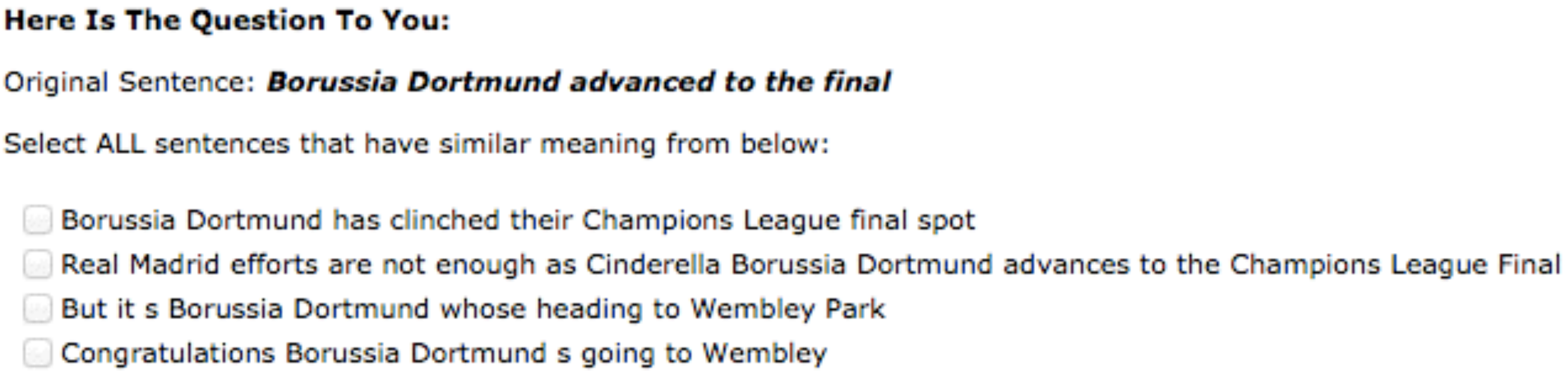
Annotation

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

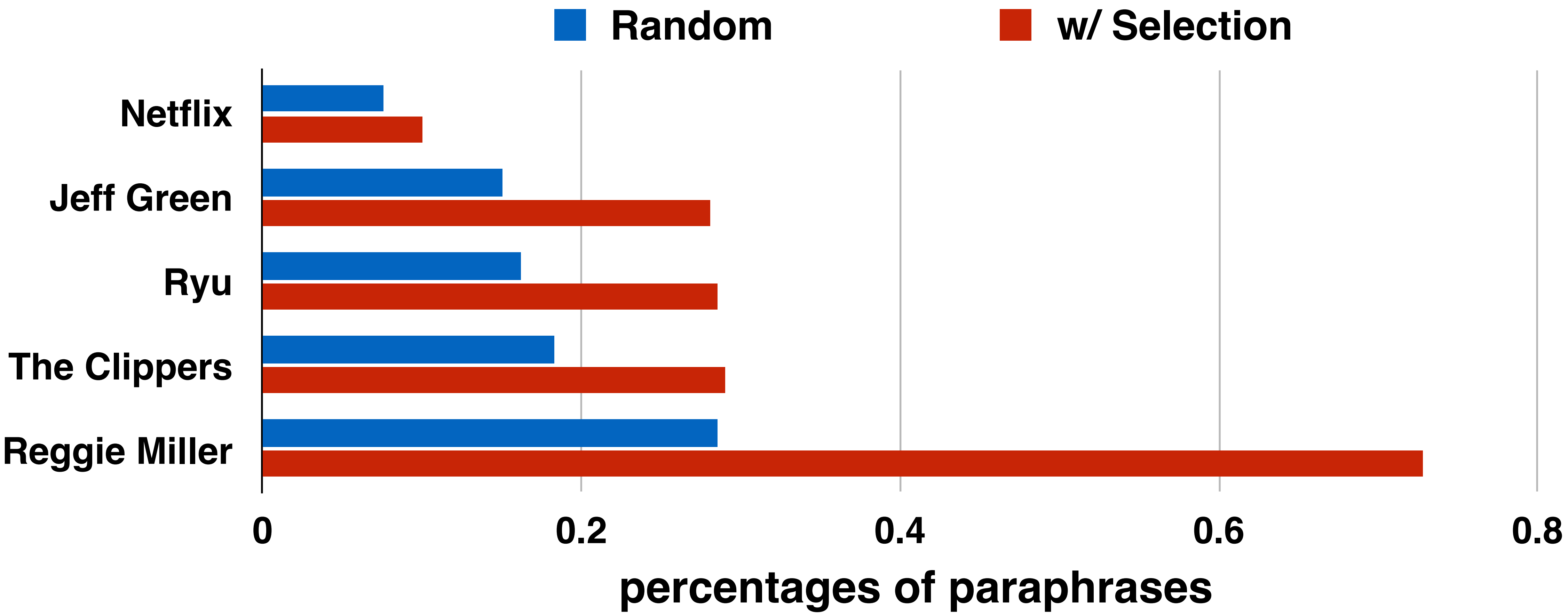
Wei Xu. "Data-driven Approaches for Paraphrasing Across Language Variations" PhD Thesis, New York University. (2014)

Sentence Selection

SumBasic Algorithm

8% → 16%

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

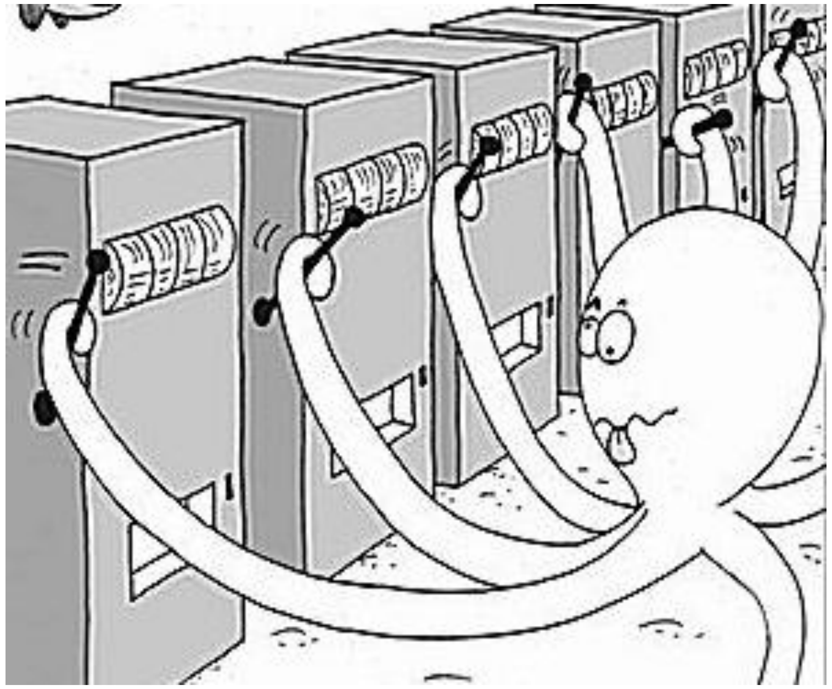


Wei Xu. "Data-driven Approaches for Paraphrasing Across Language Variations" PhD Thesis, New York University. (2014)

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

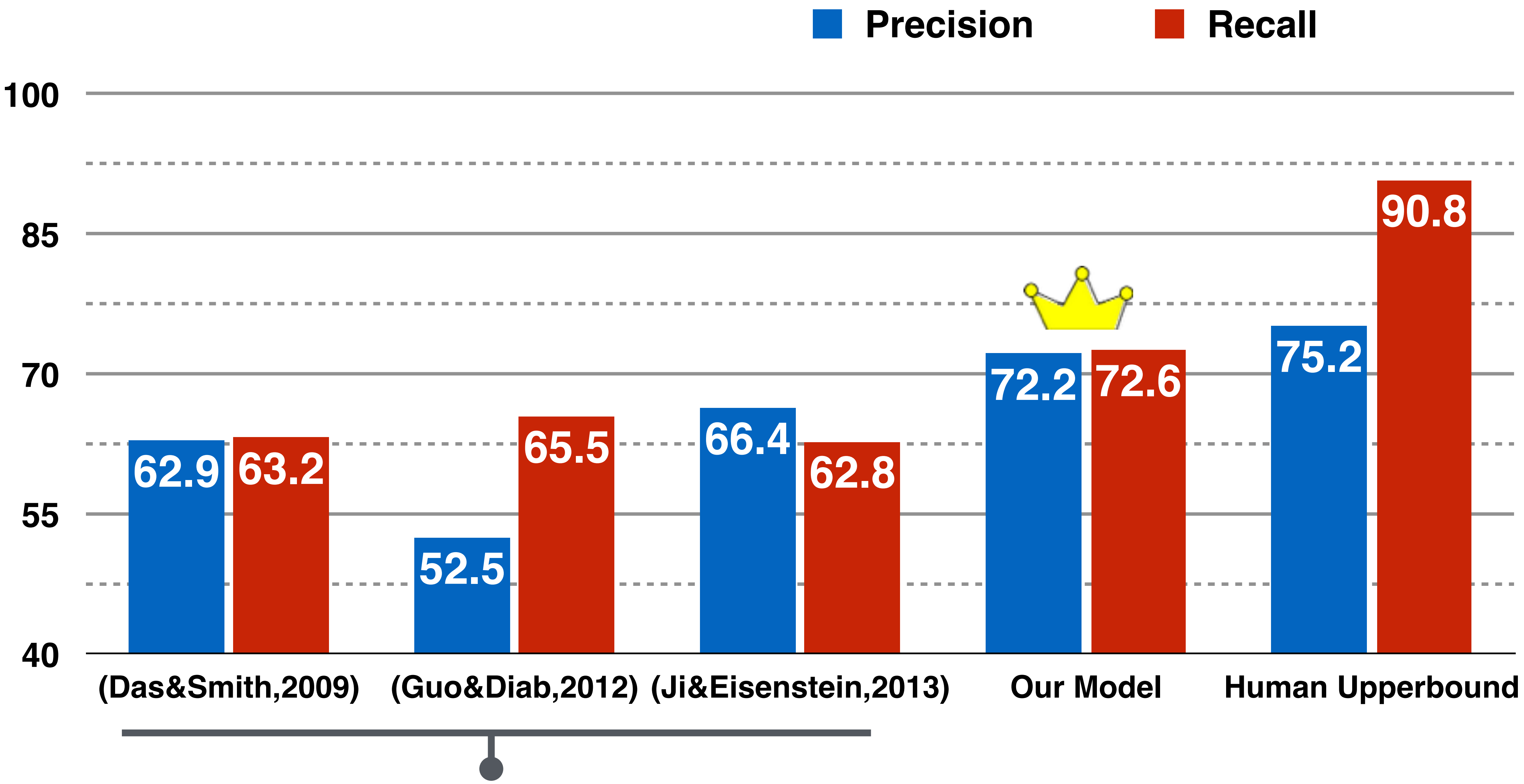
1/3 paraphrase, 2/3 non-paraphrase (very balanced)

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

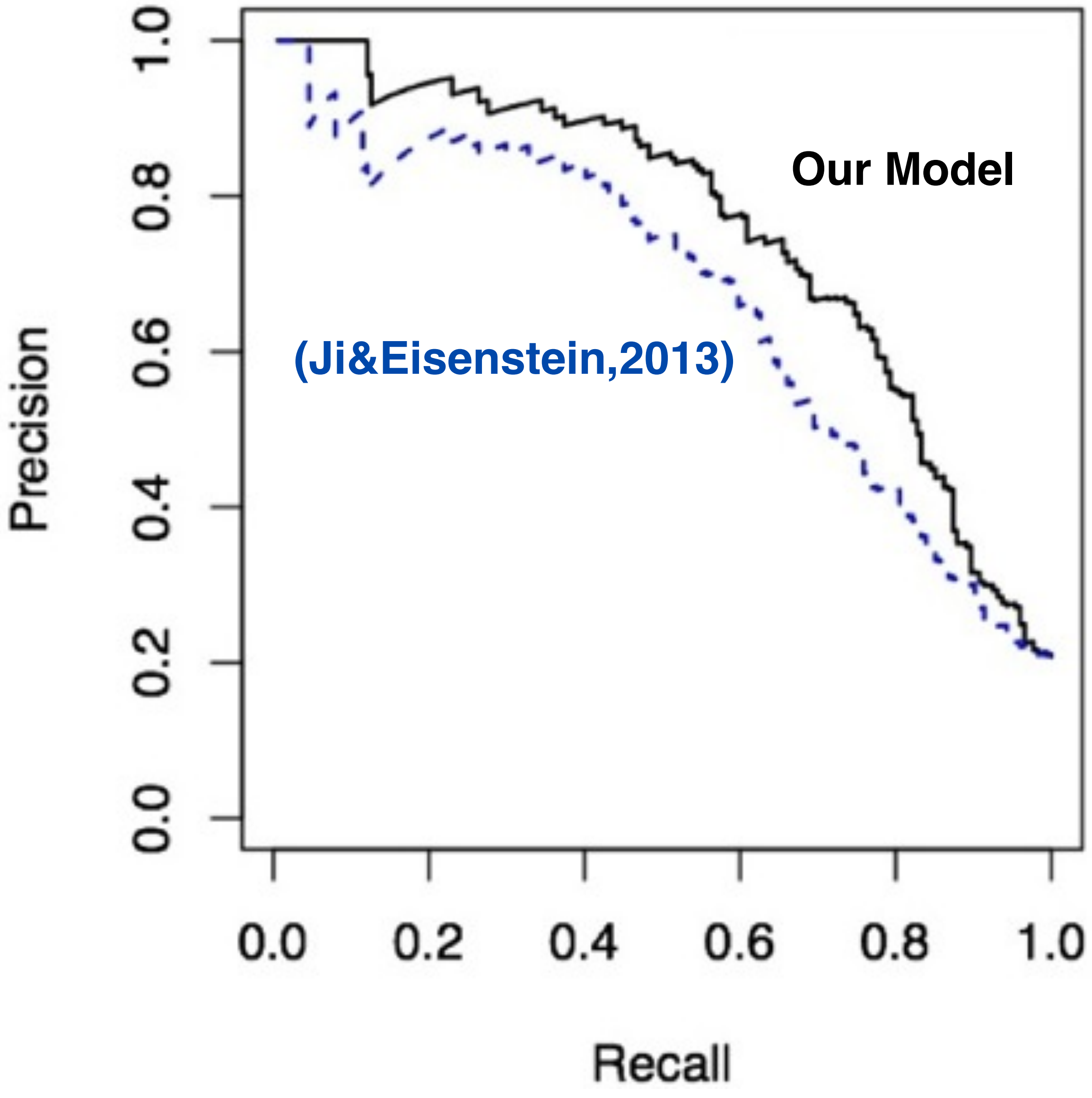
Performance

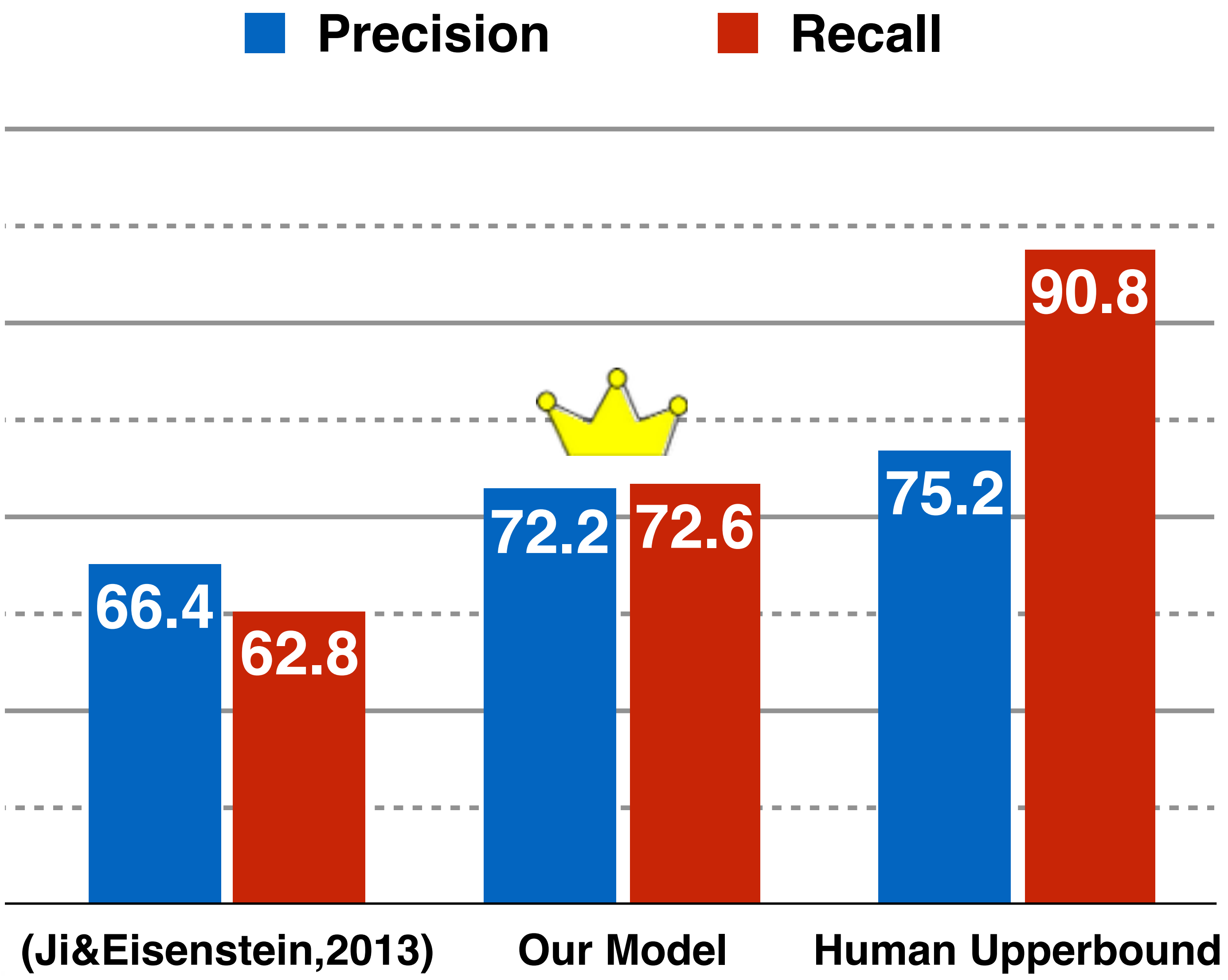


state-of-the-art of paraphrase identification

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)

Impact

SemEval 2015 shared task on “Paraphrase in Twitter”
19 + 1 teams participated

100+ research groups
have requested the data since Nov 2014

paraphrase identification (0 or 1)	rank 1	our model
semantic similarity (0 ~ 1)	rank 4	

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

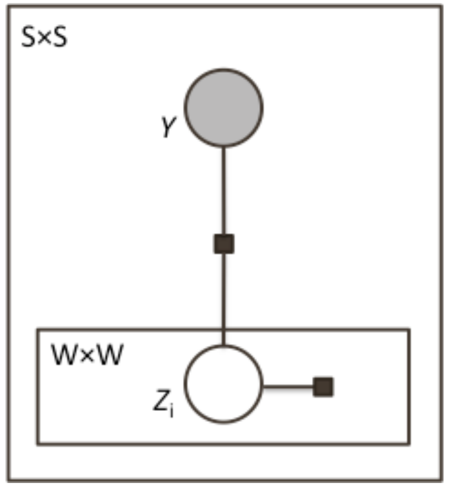
Innovations

That boy Brook Lopez with a deep **3**

brook lopez hit a **3**

Yes!

Multi-instance Learning Paraphrase Model (MultiP)



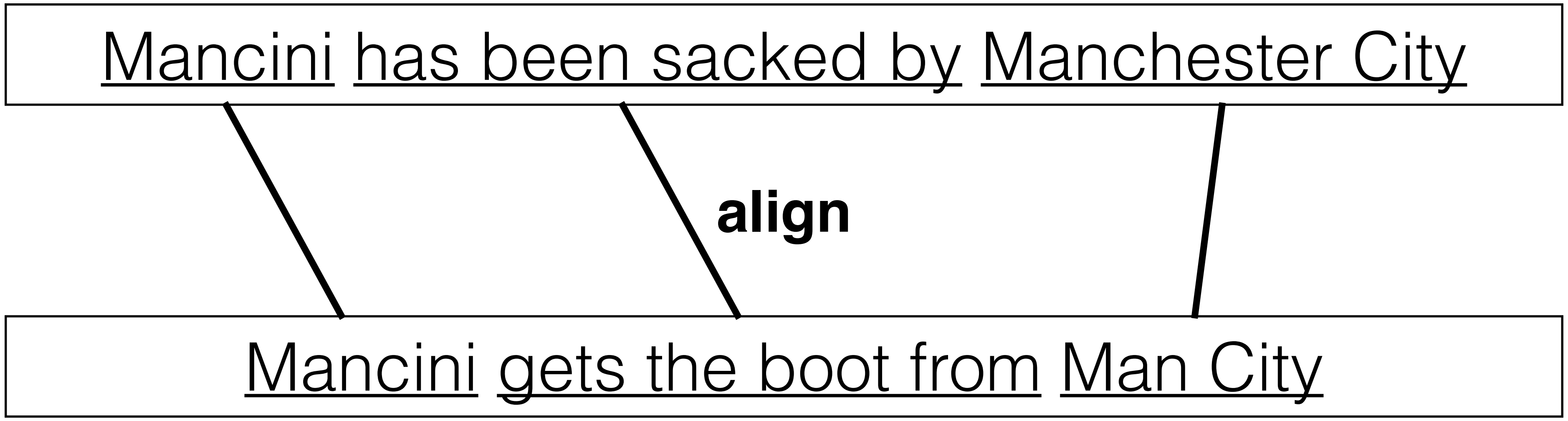
- Twitter’s big data stream
- potential beyond Twitter and English
- joint sentence-word alignment
- extensible latent variable model

(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)

Generate Paraphrases

Extract Phrasal Paraphrases



Mancini has been sacked by Manchester City

Mancini gets the boot from Man City

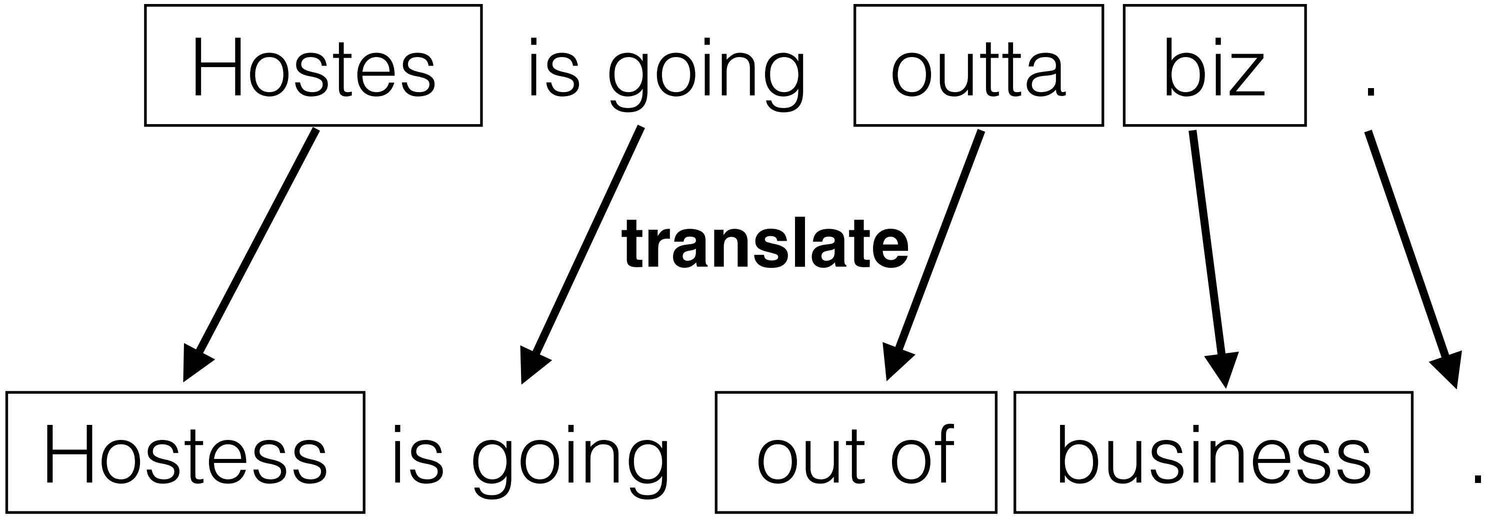
Wei Xu, Alan Ritter, Ralph Grishman. "Gathering and Generating Paraphrases from Twitter with Application to Normalization"
In BUCC (2013)

Extract Phrasal Paraphrases

has been sacked by	gets the boot from
manchester city	man city
4	for
4	four
outta	out of
hostes	hostess

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

Text-to-text Generation



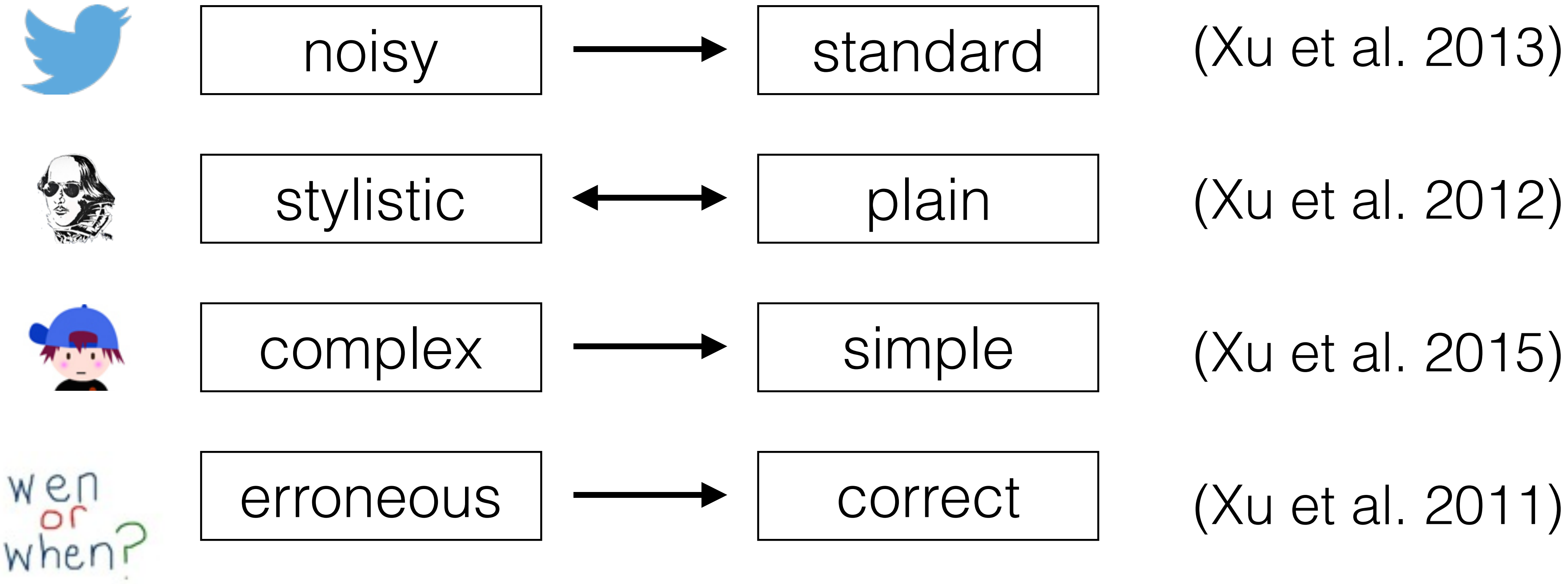
Wei Xu, Alan Ritter, Ralph Grishman. "Gathering and Generating Paraphrases from Twitter with Application to Normalization"
In BUCC (2013)

Statistical Machine Translation

	Bilingual	(Paraphrase =) Monolingual
studied	a lot	more recently
naturally available parallel text	more	less
sensitive to error	less	more
objective	straightforward	sophisticated
has standard evaluation	yes	not quite yet

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

Text-to-text Generation



noisy → standard (Xu et al. 2013)

stylistic ↔ plain (Xu et al. 2012)

complex → simple (Xu et al. 2015)

erroneous → correct (Xu et al. 2011)

and more (future work) …

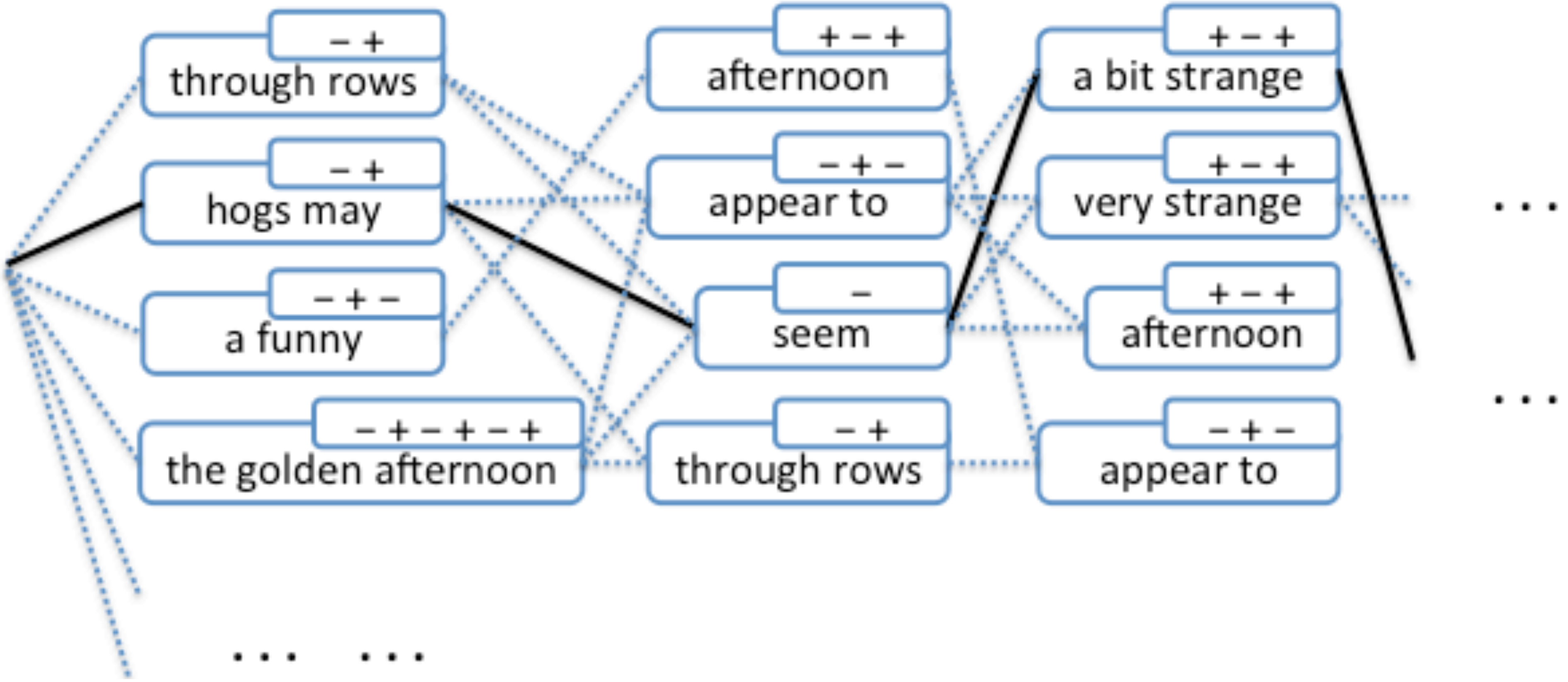
Wei Xu. “Data-driven Approaches for Paraphrasing Across Language Variations” PhD Thesis, New York University. (2014)

Prose to Sonnet

Wandering through rows of stalls examining workhorses and prize hogs may seem to … have been a strange way for a scientist to spend an afternoon, but there was a certain logic to it.

hogs may seem a bit strange through rows of stalls

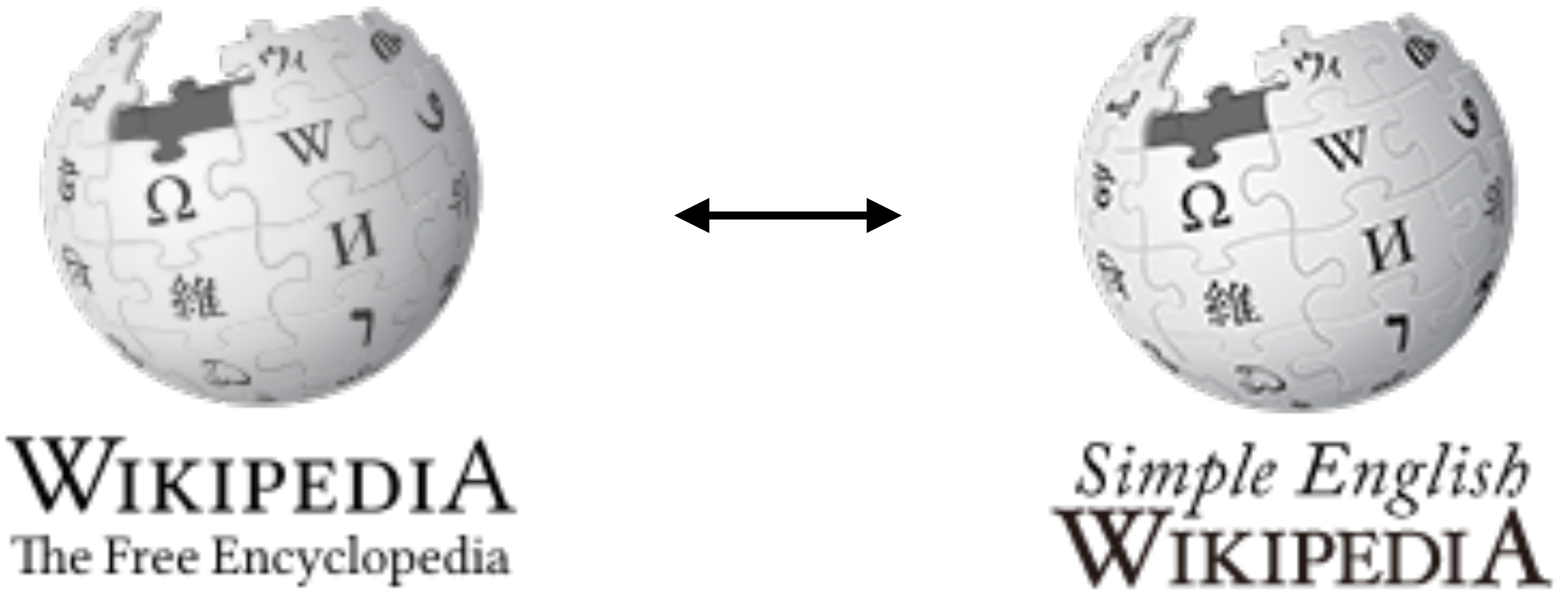
[Rhyme]
balls
falls
installs
walls
…



Quanze Chen, Chenyang Lei, Wei Xu, Ellie Pavlick and Chris Callison-Burch. "Poetry of the Crowd: A Human Computation Algorithm to Convert Prose into Rhyming Verse" In AAAI's HCOMP (2012)

Text Simplification

state-of-the-art (since 2010)

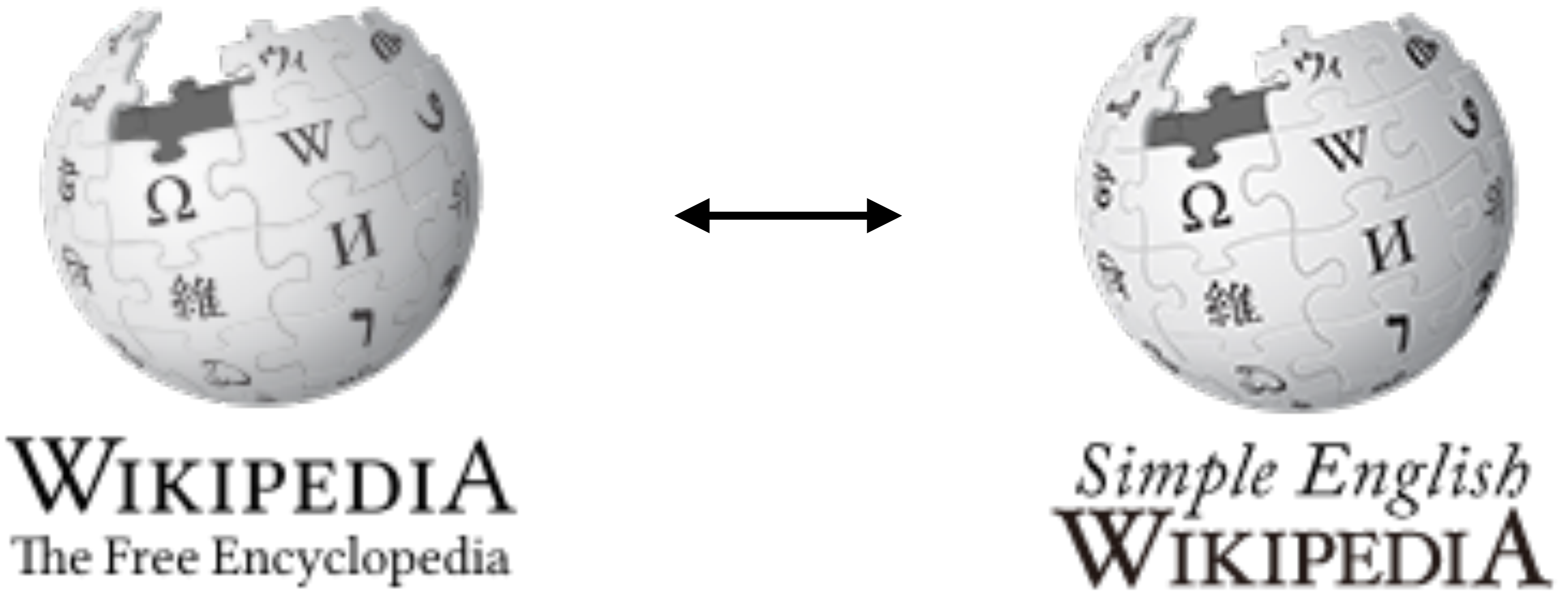


NSF EAGER: “Simplification as Machine Translation” (2014 ~ 2015)

Text Simplification

state-of-the-art (since 2010)

is suboptimal !



is not all that simple

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

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

NEW

Main Contributions

- **Jointly model word-sentence via latent variables**
- **Use Twitter as a powerful paraphrase resource**
- **Systemize a framework for language generation**
- **Right the direction of text simplification research**

all code and data are available on my homepage: http://www.cis.upenn.edu/~xwe/

The Ideal

You've suffered a contusion to the soft tissue below the fourth thoracic vertebra, exacerbating the proximal sternum.
www.CartoonStock.com
search ID: jlvn820
© LoveA711

Translation: "You have a bruised rib."

Collaborators

Chris Callison-Burch	UPenn
Ralph Grishman	NYU
Bill Dolan	MSR
Alan Ritter	UW / OSU
Raphael Hoffmann	UW / AI2 Incubator
Joel Tetreault	ETS / Yahoo!
Le Zhao	CMU / Google
Maria Pershina	NYU
Martin Chodorow	CUNY
Colin Cherry	NRC
Yangfeng Ji	GaTech
Ellie Pavlick	UPenn
Mingkun Gao	UPenn
Quanze Chen	UPenn

thank u 4 ur time

Thank you

thanking you

appreciate it

gratitude

thx

3x

tyvm

thanks

thank you very much

say thanks

thnx

wawwww thankkkkkkkkkkk you alottttttttttttt!

thanks a lot

am grateful