Paraphrases in Twitter

Wei Xu
Computer and Information Science
University of Pennsylvania

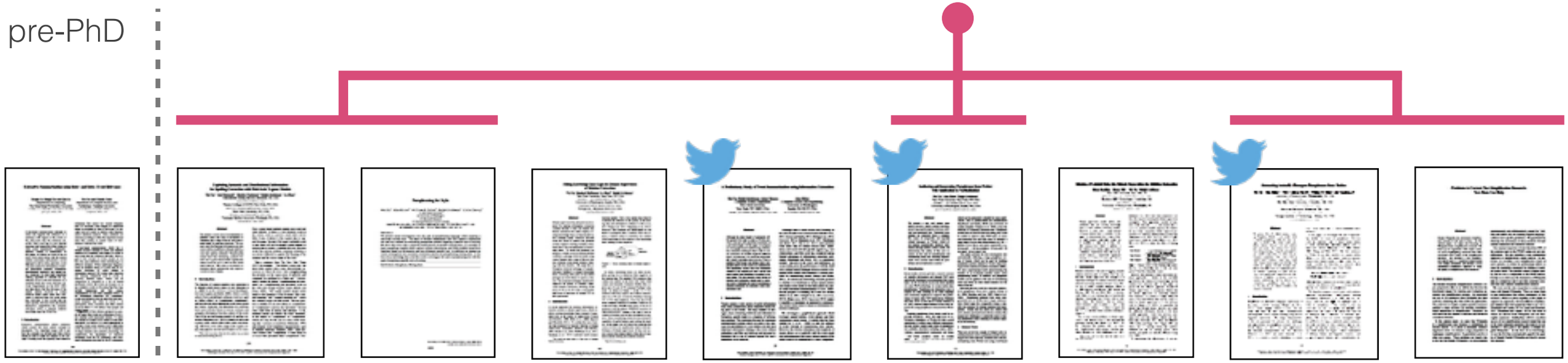
Follow @cocoweixu

Talk @ TwitterHQ Feb-6-2015

Research Overview

Learning & Generating Paraphrases

pre-PhD



ACL 06 | EMNLP 11 | COLING 12 | ACL 13 | LSAM 13 | BUCC 13 | ACL 14 | TACL 14 | In Review

Information Extraction

Paraphrase

Paraphrase

cup	**word**	mug
the king's speech	**phrase**	His Majesty's address
… the forced resignation of the CEO of Boeing, Harry Stonecipher, for …	**sentence**	Boeing Co. Monday said it fired Chief Executive Harry Stonecipher, after …

Application

Information Extraction

end_job (Harry Stonecipher, Boeing)
title (Harry Stonecipher, CEO)

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

Boeing Co. Monday said it fired Chief Executive Harry Stonecipher, after …

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?

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

Boeing Co. Monday said it fired Chief Executive Harry Stonecipher, after …

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.

Wei Xu, Chris Callison-Burch. “Problems in Current Text Simplification Research: New Data Can Help” In Submission (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 for paraphrase identification, extraction, generation and various applications.

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

Previous Work

only a few hundreds news agencies
report big events
in 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

Twitter as a powerful resource

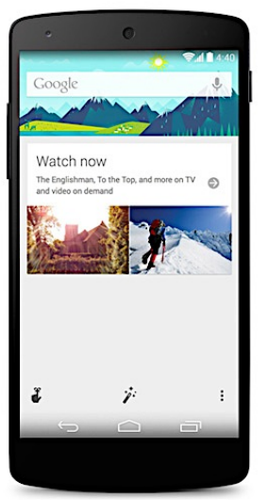
thousands of users
talk about both big and micro events
in formal, informal, erroneous languages

Enables new applications

Human-computer Interaction

who want to get a beer?

Apple Siri



Google Now

Windows Cortana

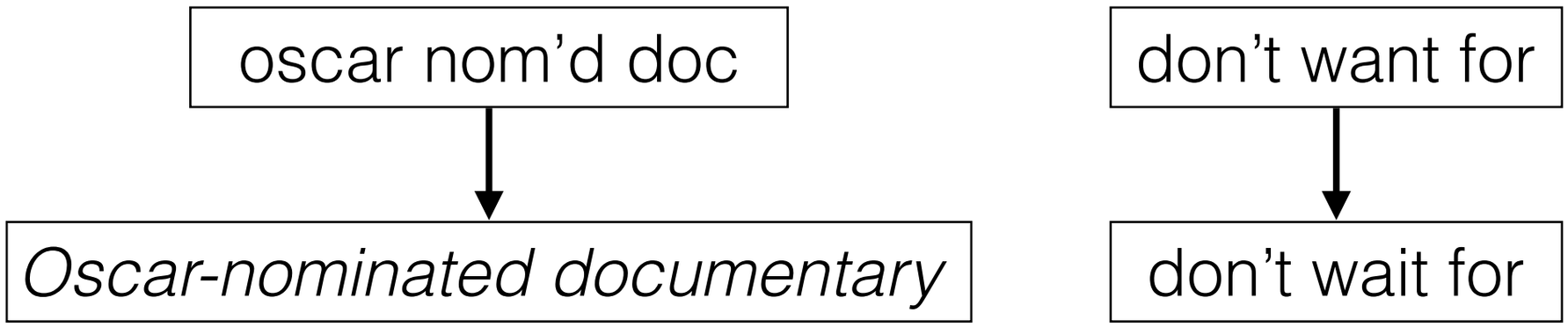
- wants to get a beer?
- someone to get a beer?
- who wants to go get a beer?
- trying to get a beer?
- who wants to buy a beer?
- who wants to get a drink?
- who else wants to get a beer?

… (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

Noisy Text Normalization



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

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 and Bulls game is bananas

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

Trilogy of Paraphrase

1

Trilogy of Paraphrase

#1 identify parallel sentences

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!

Design a model

Train it by data

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 Lopze with a deep **3**

brook lopez hit a **3**

A 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!

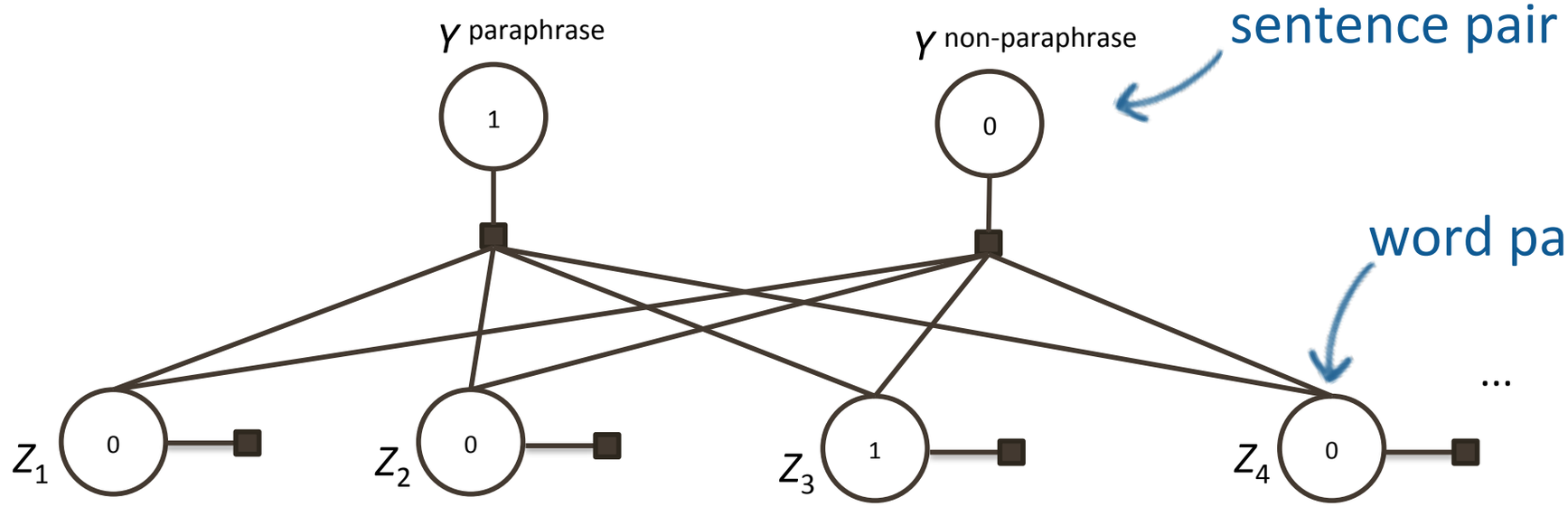
Solution:

a discriminative model using features at word-level

Multi-instance Learning Paraphrase Model

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

Teo is the little **new** Junior Seau

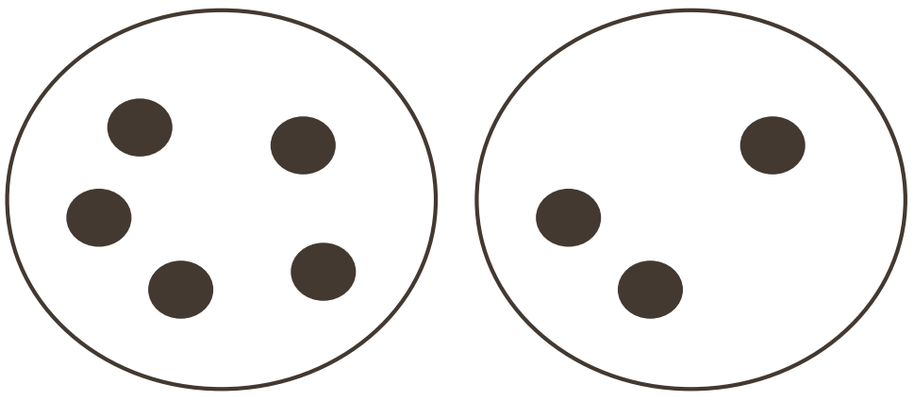


Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan and Yangfeng Ji. "Gathering and Generating Paraphrases from Twitter with Application to Normalization" In Transactions of the Association for Computational Linguistics (2014)

[Mini Tutorial]
Multi-instance Learning

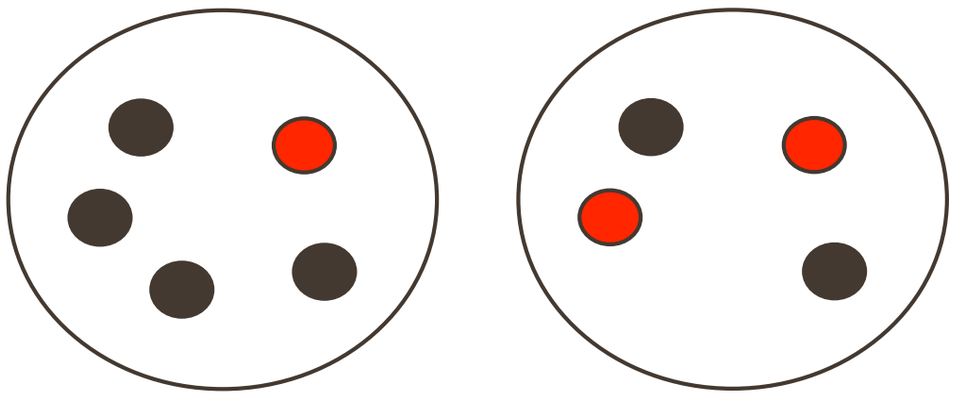
Instead of labels on each individual instance,
the learner only observes labels on bags of instances.

Negative Bags



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

Positive Bags

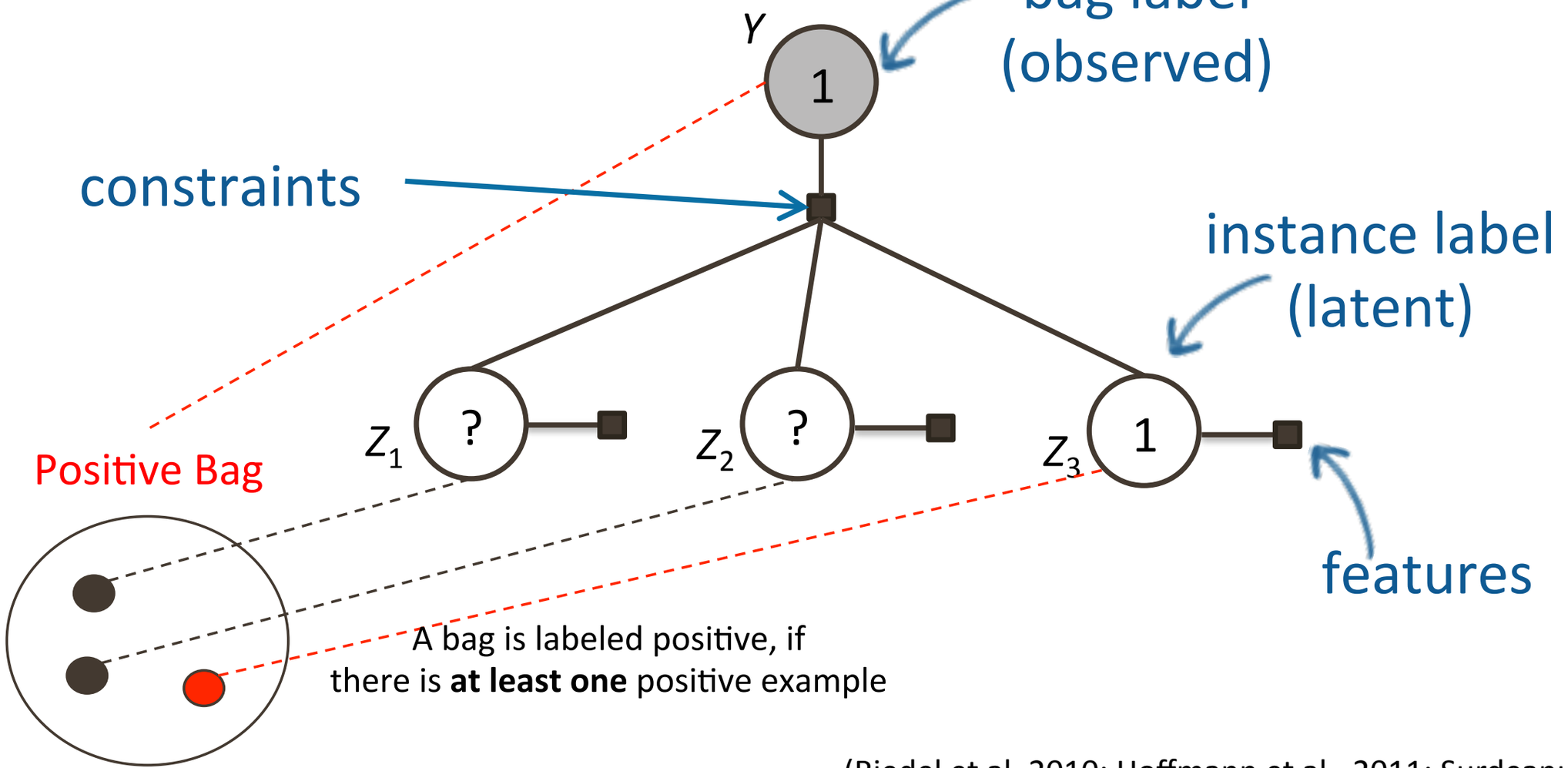


A bag is labeled positive, if
there is **at least one** positive example

(Dietterich et al., 1997)

[Mini Tutorial]
Multi-instance Learning

Latent Variable Model

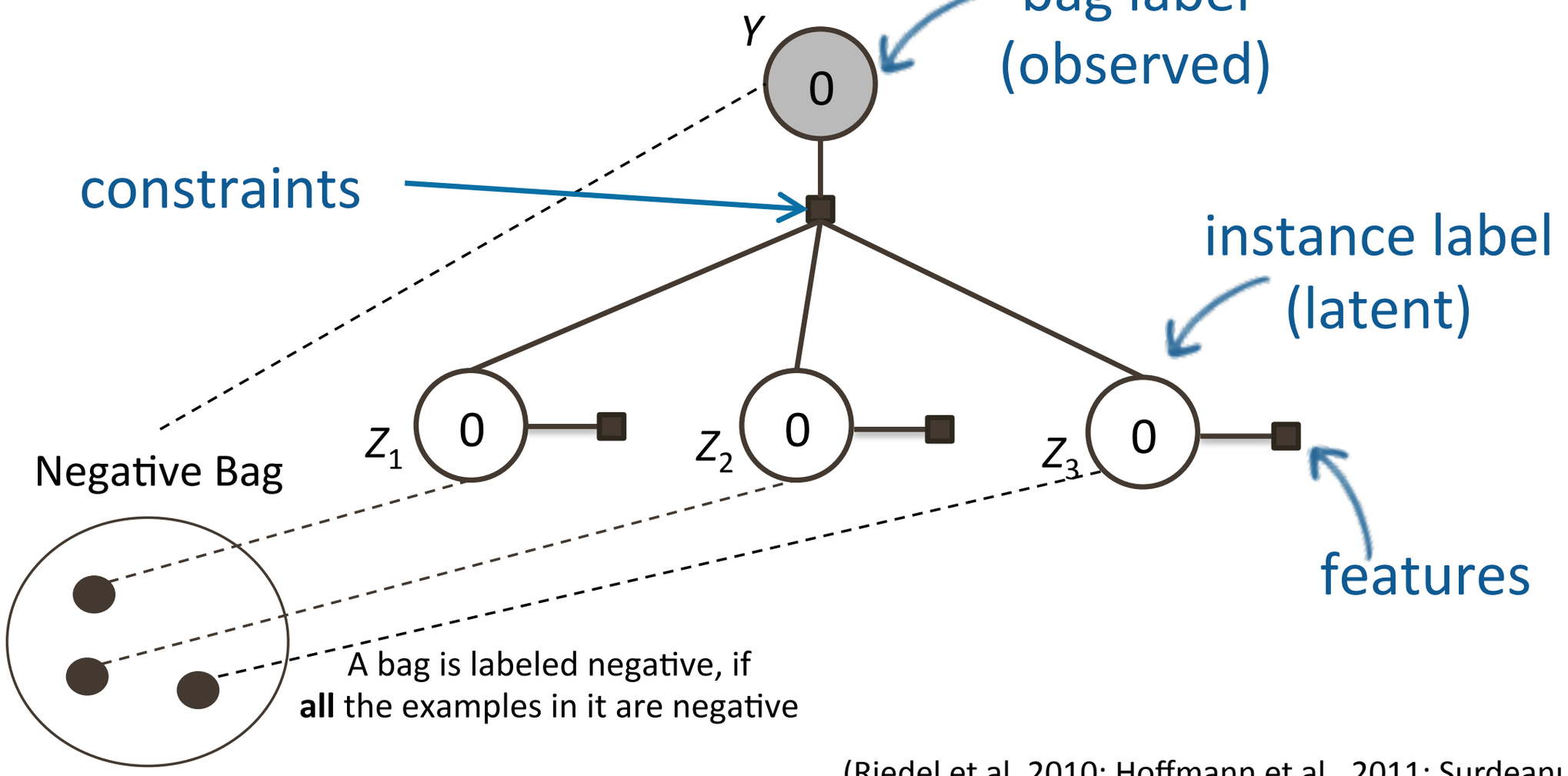


(Riedel et al. 2010; Hoffmann et al., 2011; Surdeanu et al. 2012)

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

[Mini Tutorial]
Multi-instance Learning

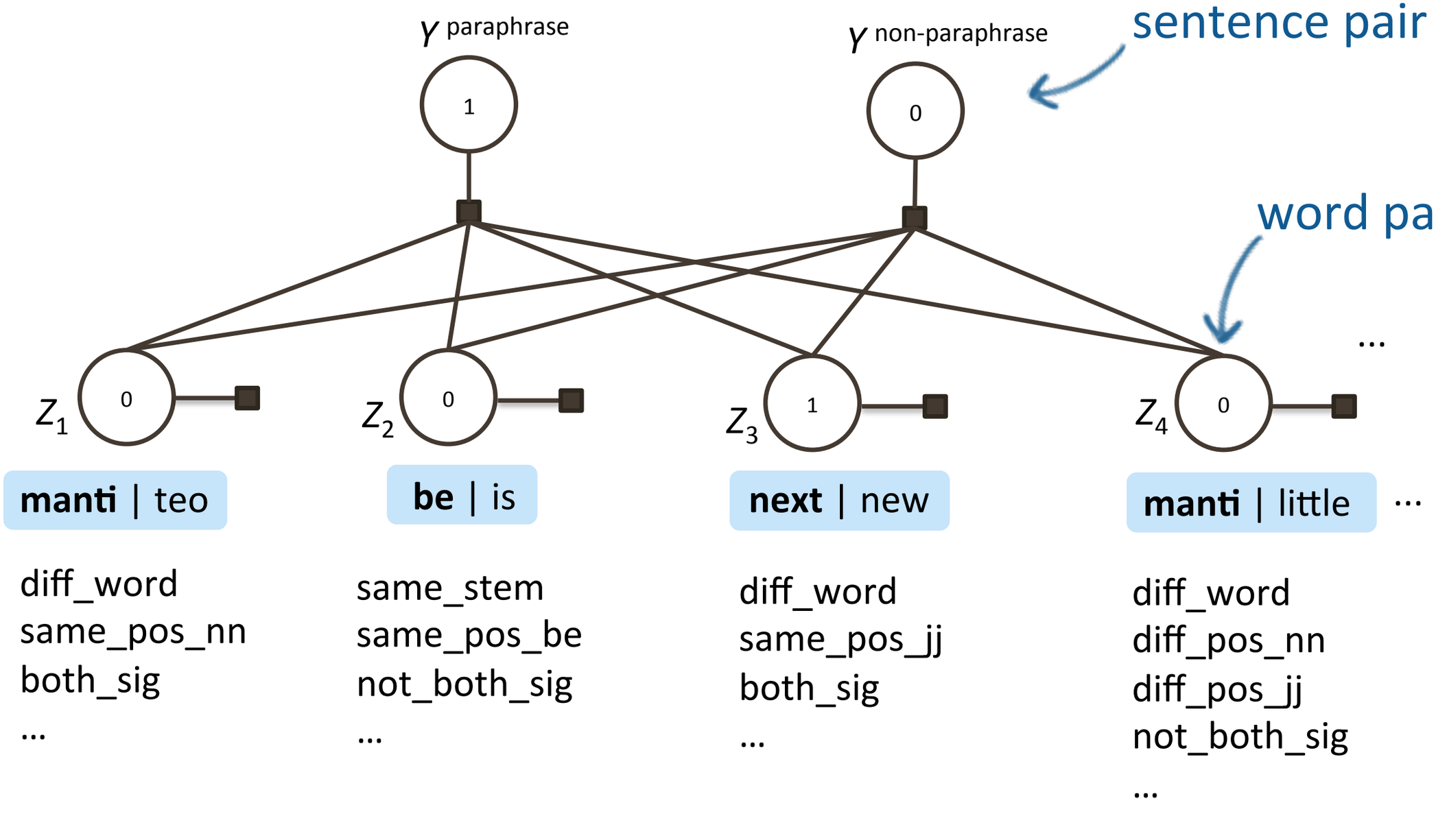
Latent Variable Model



(Riedel et al. 2010; Hoffmann et al., 2011; Surdeanu et al. 2012)

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

[Recap] Multi-instance Learning Paraphrase Model



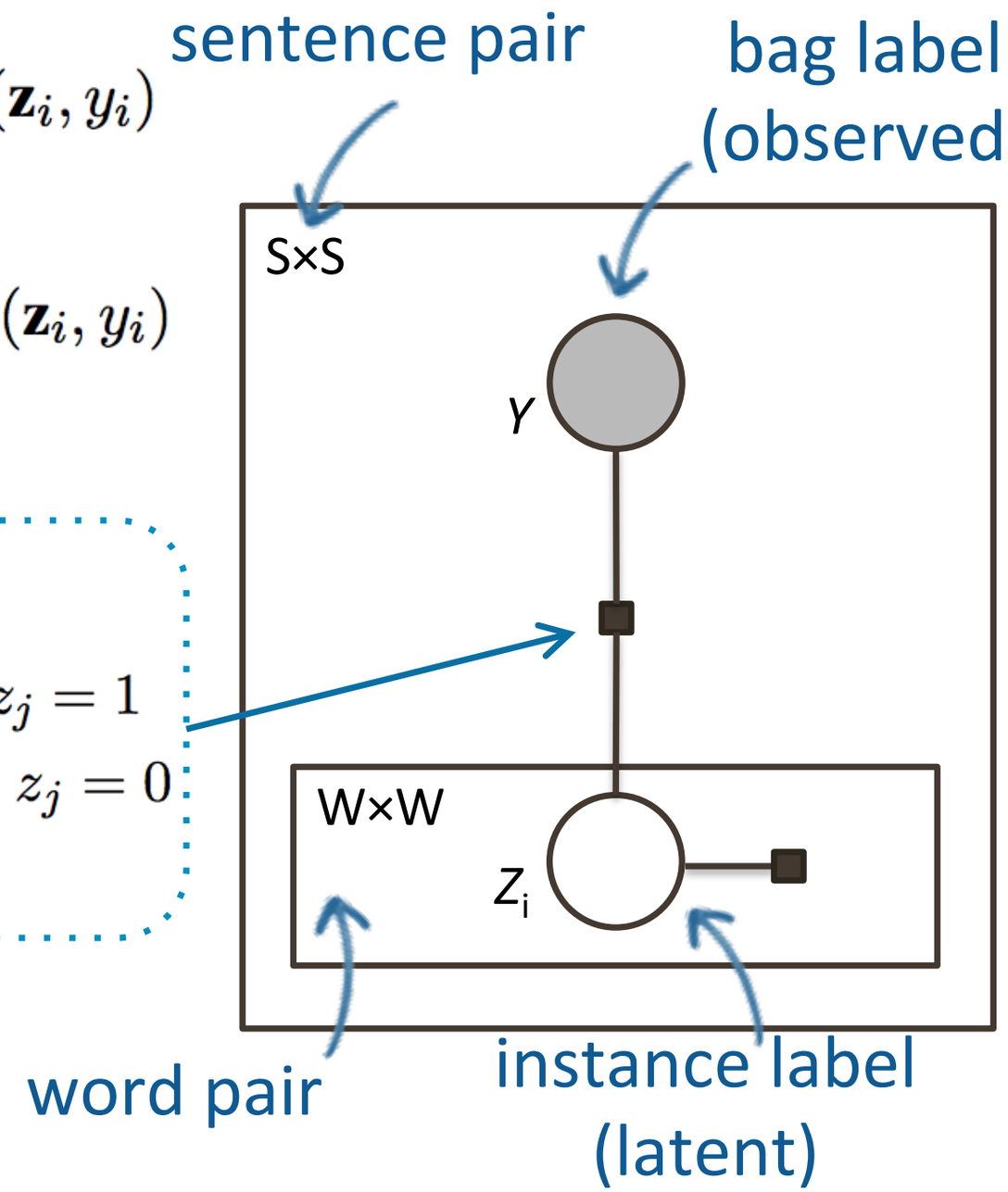
Joint word-sentence inference

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

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

deterministic OR

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



Learning algorithm

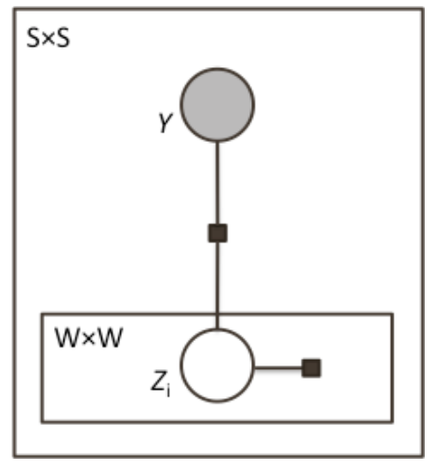
Objective

parameters

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

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

all sentence pairs in training set



Update

features

word-level only

$$\frac{\partial \log P(\mathbf{y}|\mathbf{w};\theta)}{\partial\theta} = \mathbf{E}_{P(\mathbf{z}|\mathbf{w},\mathbf{y};\theta)}\Big(\sum_i f(\mathbf{z}_i, \mathbf{w}_i)\Big) - \mathbf{E}_{P(\mathbf{z},\mathbf{y}|\mathbf{w};\theta)}\Big(\sum_i f(\mathbf{z}_i, \mathbf{w}_i)\Big)$$

conditioned on sentence-level label

$$\approx \sum_i f(\mathbf{z}_i^*, \mathbf{w}_i) - \sum_i f(\mathbf{z}_i', \mathbf{w}_i)$$

[Recap] Task

#1 identify parallel sentences

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!

[Recap] Model

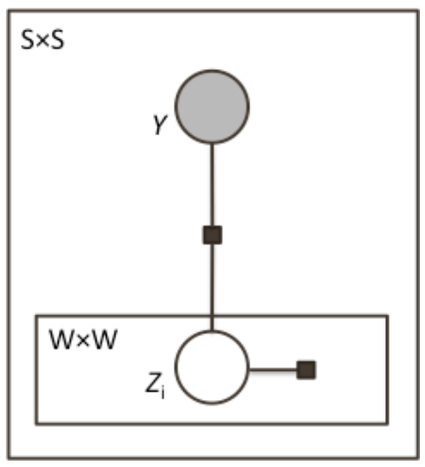
observation

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

brook lopez hit a **3**

Yes!

Multi-instance Learning Paraphrase Model (MultiP)



- big data stream
- principled latent variable model
- flexible extensible framework
- no word-level annotation needed

Training data

Training data

Crowdsourcing

(Courtesy: The Sheep Market)

Training data

Crowdsourcing

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

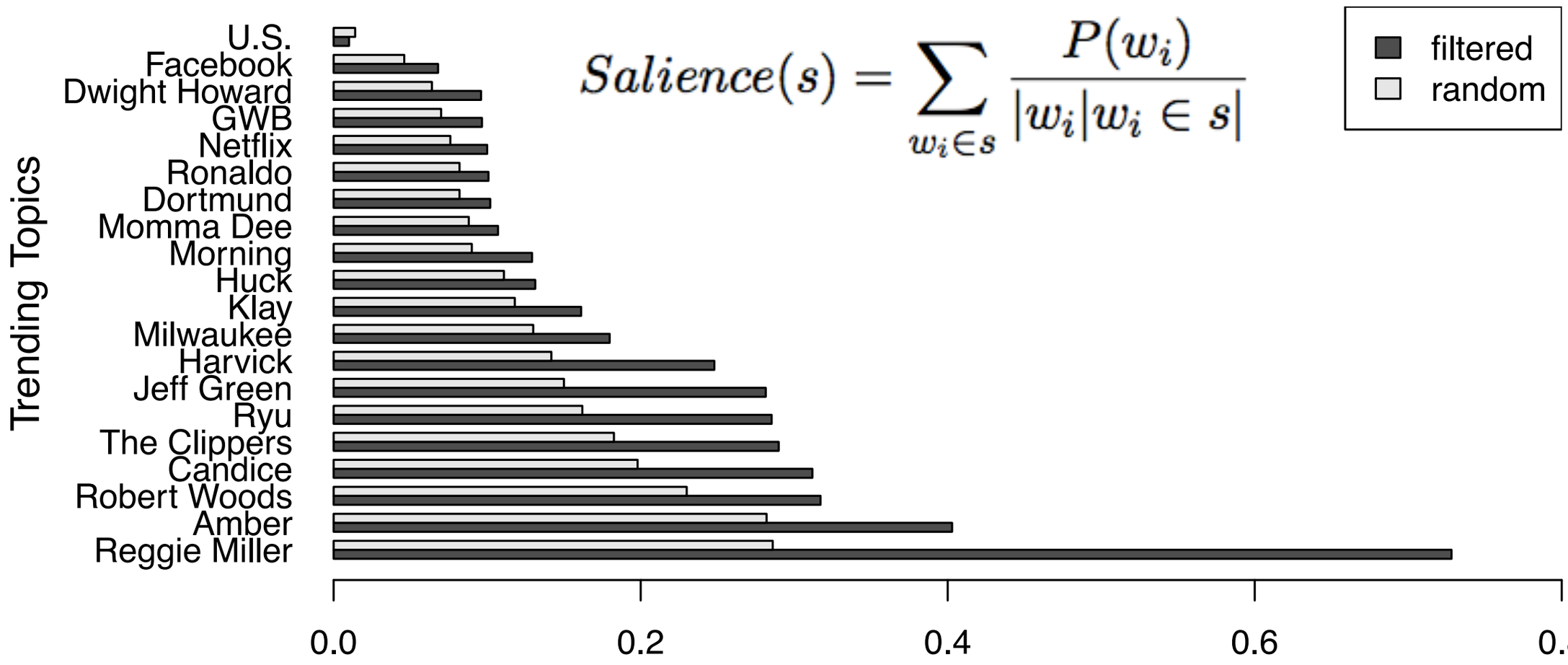
only **8%** sentence pairs about the same topic have similar meaning

hurt both quantity and quality

non-experts lower their bars

Sentence Selection

8% → 16%



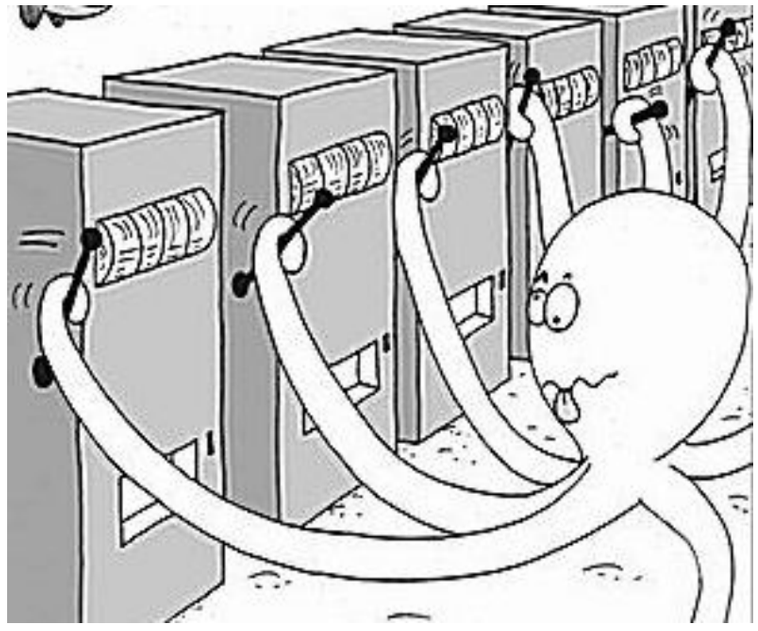
$$Salience(s) = \sum_{w_i \in s} \frac{P(w_i)}{|w_i|w_i \in s|}$$

filtered
random

Topic Selection

Multi-Armed Bandits

16% → 34%



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

$$\text{s.t.} \sum_{j \in \{i | r_i(t_1) > 0\}} r_i(t_0) \leq (1 - \epsilon) B, \forall i : 0 \leq r_i(t_1) \leq l - r_i(t_0).$$

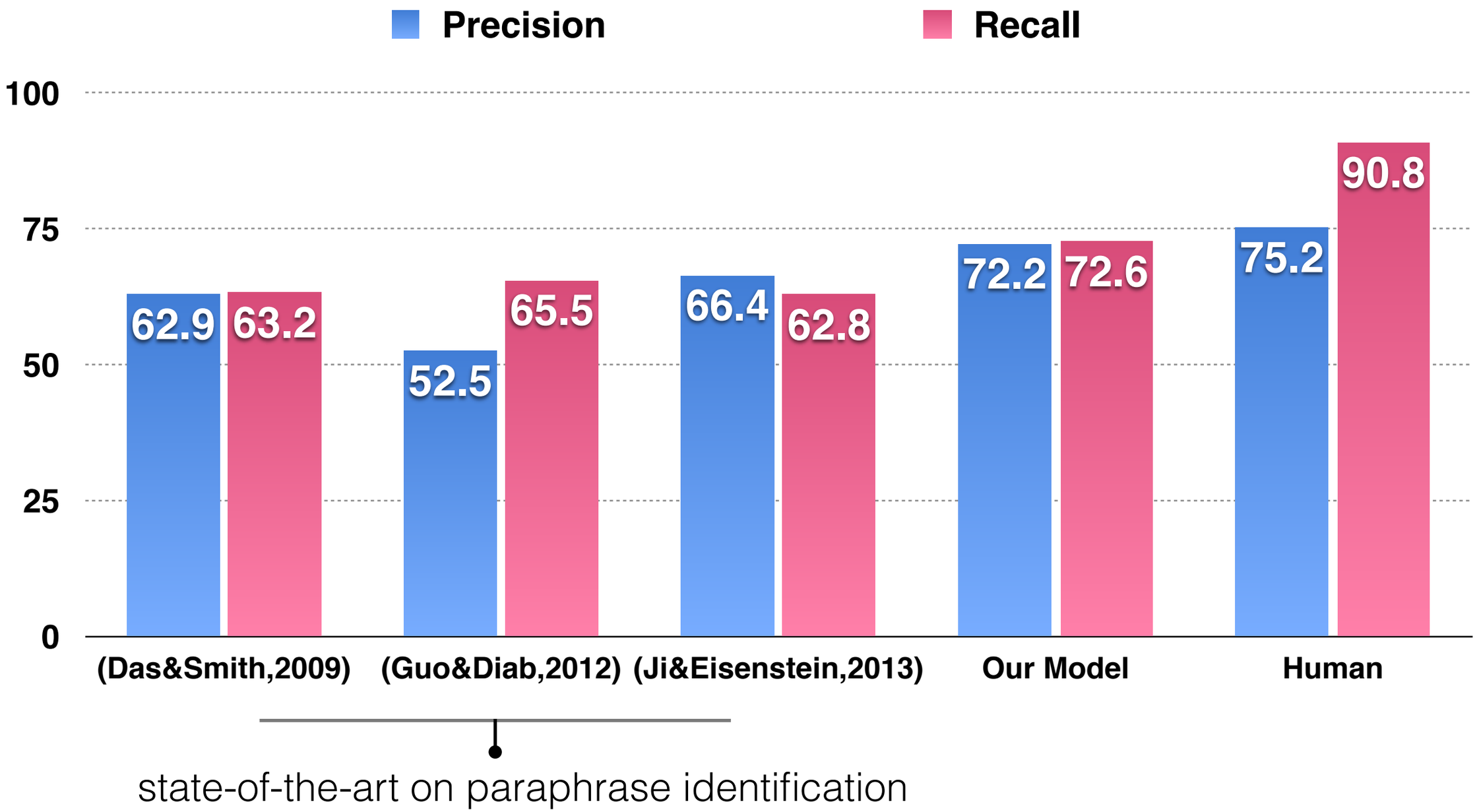
Training data

18,762 sentence pairs labeled
cost only $200
1/3 paraphrase, 2/3 non-paraphrase (very balanced)

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

Experiment

Experiment


Precision
Recall
100
75
50
25
0
62.9
63.2
52.5
65.5
66.4
62.8
72.2
72.6
75.2
90.8
(Das&Smith,2009)
(Guo&Diab,2012)
(Ji&Eisenstein,2013)
Our Model
Human
state-of-the-art on paraphrase identification

Semantic Evaluation 2015

based on our Twitter paraphrase dataset
19 + 1 teams participated
some used neural networks

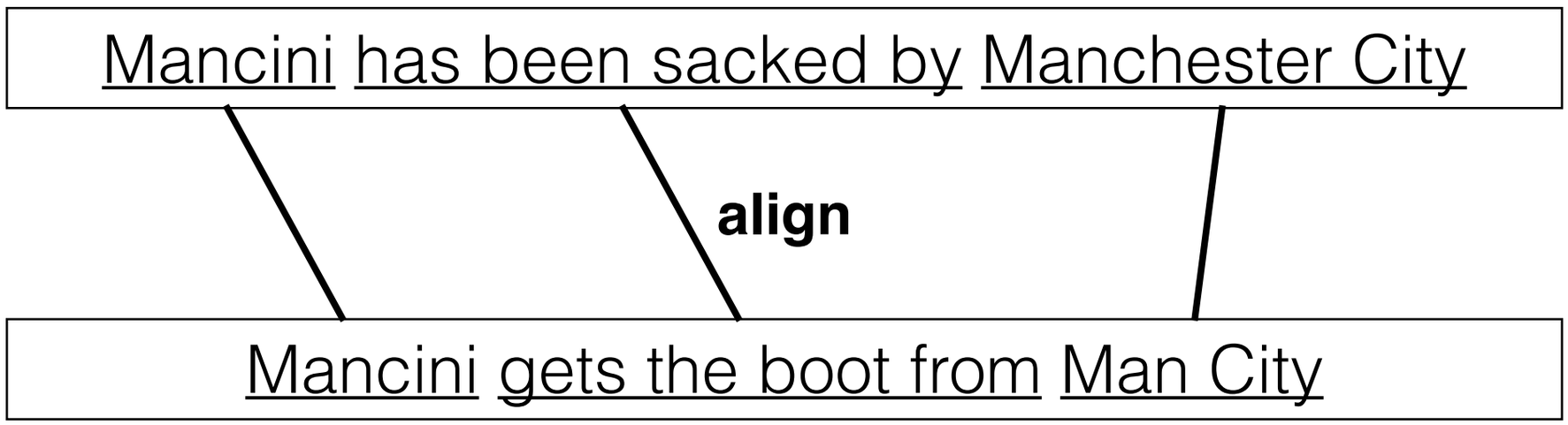
paraphrase identification (0 or 1)	rank 1 
semantic similarity (0 ~ 1)	rank 4

Trilogy of Paraphrase

2 & 3

Trilogy of Paraphrase

#2 extract lexical/phrasal paraphrase



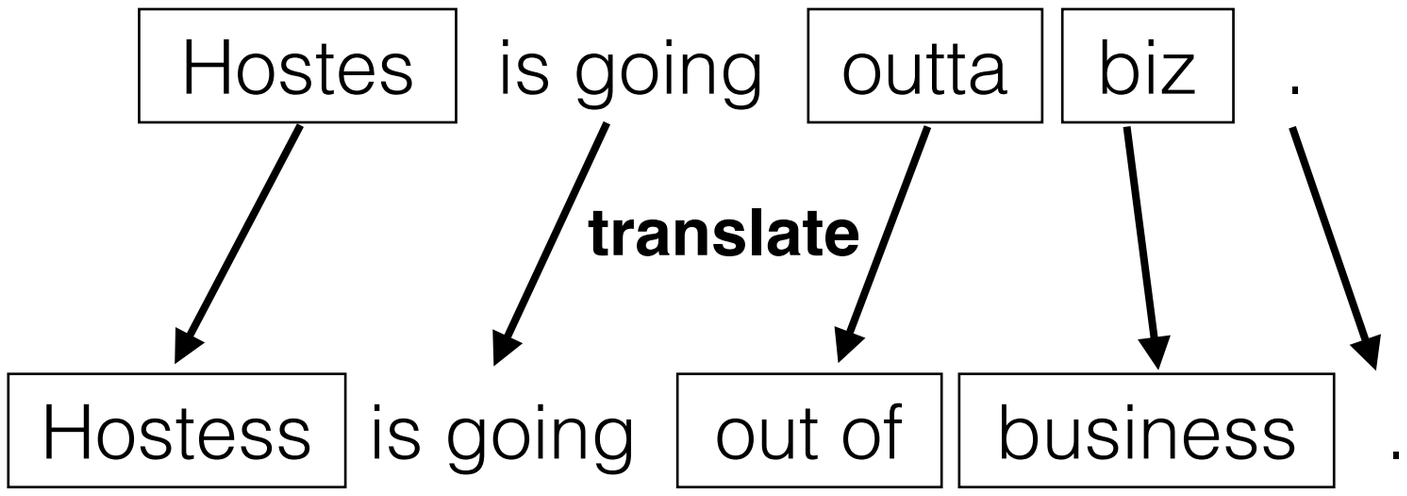
Trilogy of Paraphrase

#2 extract lexical/phrasal paraphrase

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

Trilogy of Paraphrase

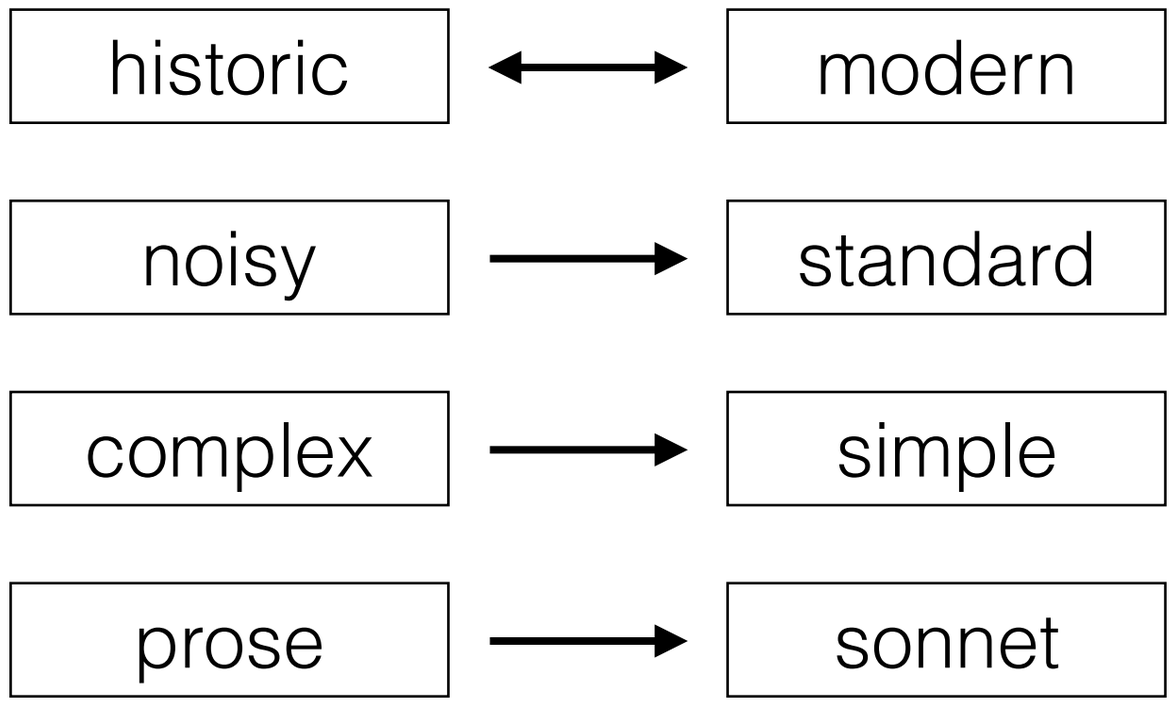
#3 text-to-text generation



Statistical Machine Translation

	Bilingual	**Monolingual**
studied	a lot	getting more
sensitive to error	less	more
naturally available parallel text	more	less
objective	simple	flavored
has standard evaluation	yes	no

Text-to-text Generation

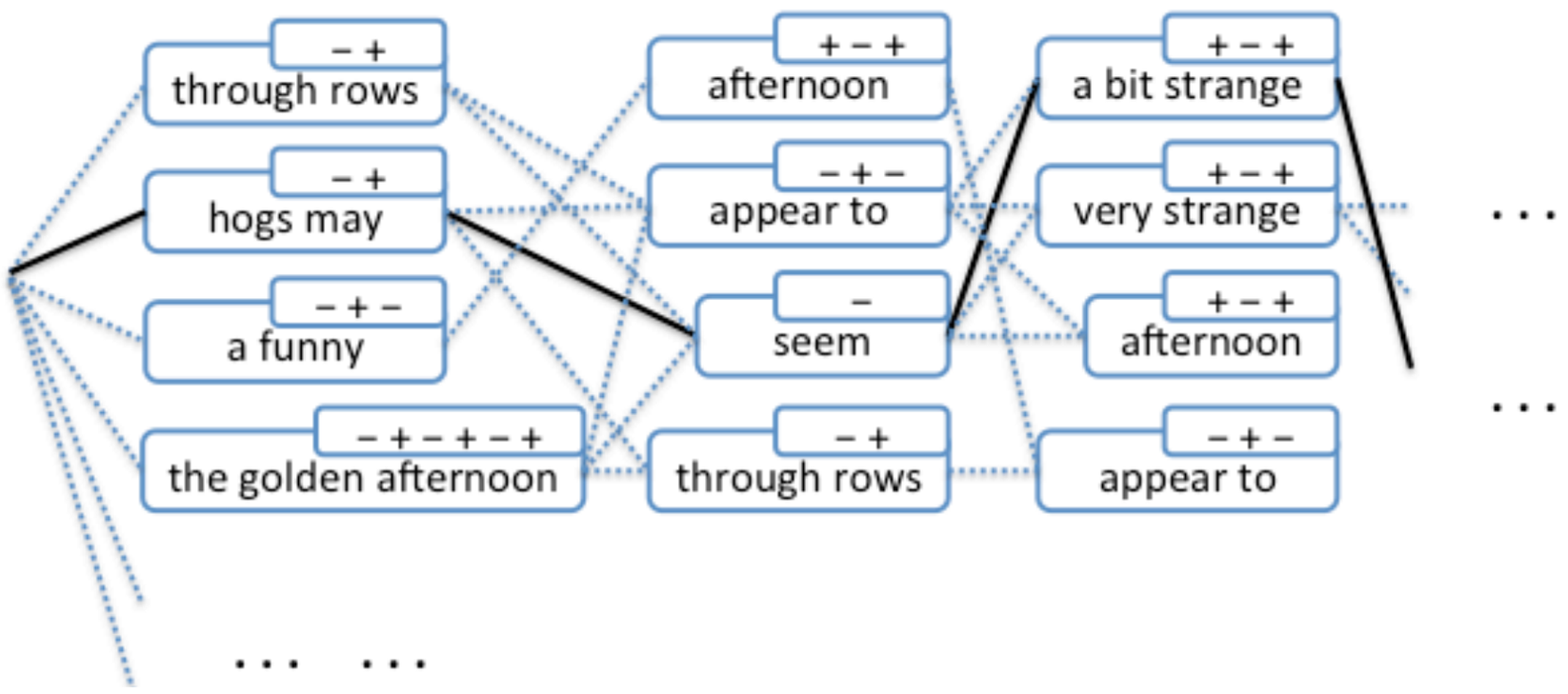


Prose to Sonnet

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



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)

The Ideal

The Ideal

Translation: "You have a bruised rib."

Collaborators

Chris Callison-Burch	UPenn
Bill Dolan	MSR
Alan Ritter	UW / OSU
Ralph Grishman	NYU
Raphael Hoffmann	UW / AI2 Incubator
Joel Tetreault	ETS / Yahoo!
Le Zhao	CMU / Google
Colin Cherry	NRC
Yangfeng Ji	GaTech

thank u 4 ur time

Thank you

thanking you

appreciate it

gratitude

thx

3x

tyvm

thanks

say thanks

thank you very much

thnx

wawwww thankkkkkkkkkk you alottttttttttt!

thanks a lot

am grateful

more details on my homepage: http://www.cis.upenn.edu/~xwe/