Social Media & Text Analysis

lecture 7 - Twitter Paraphrases and Latent Variable Models

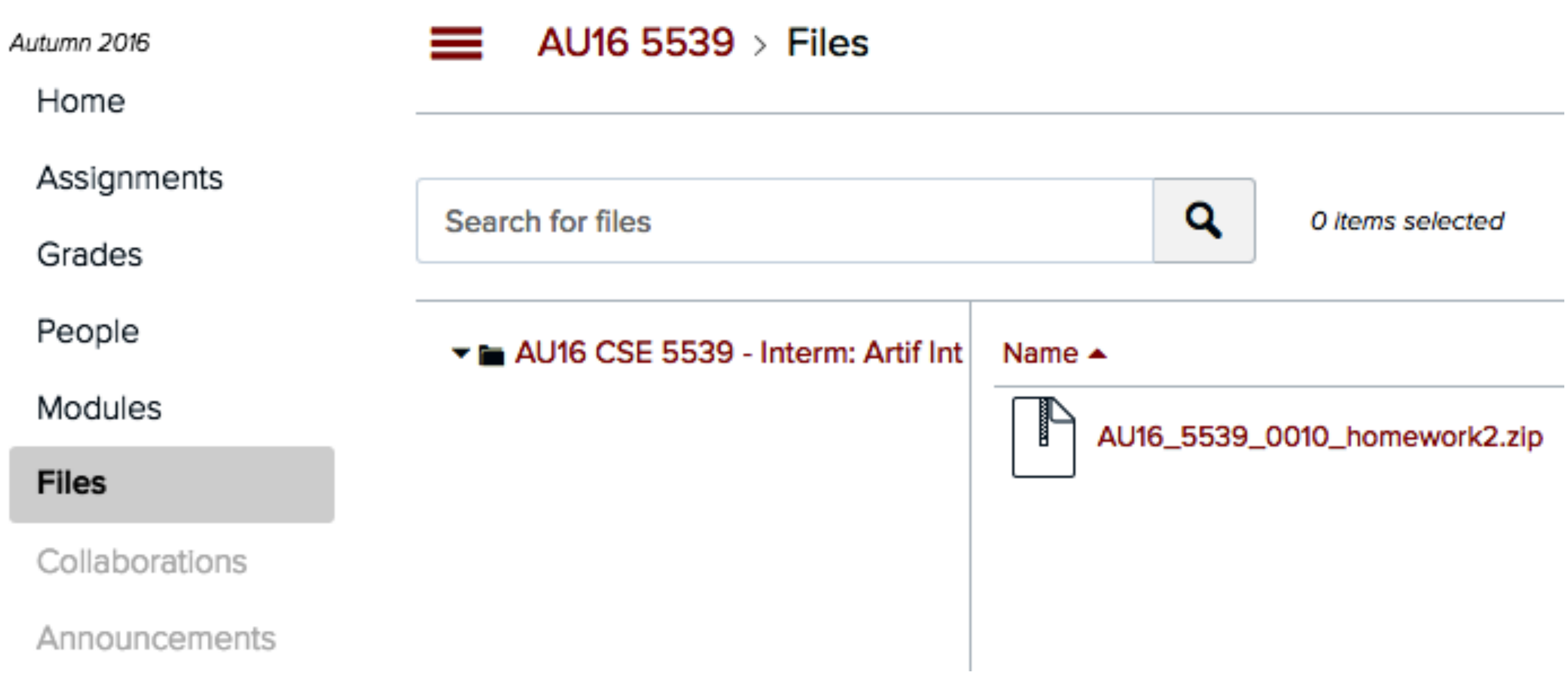
Follow @cocoweixu

CSE 5539-0010 Ohio State University

Instructor: Wei Xu

Website: socialmedia-class.org

Homework #2 is out
Due in three weeks



Collaborators

Chris Callison-Burch	UPenn
Alan Ritter	UW / OSU
Bill Dolan	MSR
Yangfeng Ji	GaTech
Jeniya Tabassum	OSU
Wuwei Lan	OSU
Ralph Grishman	NYU
Raphael Hoffmann	UW / AI2 Incubator
Joel Tetreault	ETS / Yahoo!
Le Zhao	CMU / Google
Maria Pershina	NYU
Colin Cherry	NRC
Courtney Napoles	JHU
Lyle Ungar	UPenn
Daniel Preoţiuc-Pietro	UPenn
Ellie Pavlick	UPenn
Mingkun Gao	UPenn / UIUC
Quanze Chen	UPenn / UW
Martin Chodorow	CUNY

News


Daily Mail
HIJACKING OF A VERY MIDDLE CLASS PROTEST

the guardian
'This is just the beginning'

THE TIMES
All the games
All the goals
All the action
'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

only a few hundreds news agencies
only big events
only well-edited text
(the MSR Paraphrase Corpus)

(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

average sentence length: news ≈18.6 words Twitter ≈11.9 words

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

Twitter as a powerful resource

thousands of users
talk about both big/micro events daily

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

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

Paraphrase Model

obtain sentential paraphrases automatically

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!

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

(On news data - MSR Paraphrase Corpus)

Paraphrase Identification

Algorithm	Reference	Description	Supervisi	Accurac	F1
Vector	Mihalcea et al. (2006)	cosine similarity with tf-idf weighting	unsupervised	65.40%	75.30%
ESA	Hassan (2011)	explicit semantic space	unsupervised	67.00%	79.30%
KM	Kozareva and Montoyo (2006)	combination of lexical and semantic features	supervised	76.60%	79.60%
LSA	Hassan (2011)	latent semantic space	unsupervised	68.80%	79.90%
RMLMG	Rus et al. (2008)	graph subsumption	unsupervised	70.60%	80.50%
MCS	Mihalcea et al. (2006)	combination of several word similarity measures	unsupervised	70.30%	81.30%
WTMF	Guo and Diab (2012)	latent space model for short text	unsupervised	71.51%	---
STS	Islam and Inkpen (2007)	combination of semantic and string similarity	unsupervised	72.60%	81.30%
SSA	Hassan (2011)	salient semantic space	unsupervised	72.50%	81.40%
QKC	Qiu et al. (2006)	sentence dissimilarity classification	supervised	72.00%	81.60%
ParaDetect	Zia and Wasif (2012)	PI using semantic heuristic features	supervised	74.70%	81.80%
SDS	Blacoe and Lapata (2012)	simple distributional semantic space	supervised	73.00%	82.30%
matrixJcn	Fernando and Stevenson (2008)	JCN WordNet similarity with matrix	unsupervised	74.10%	82.40%
FHS	Finch et al. (2005)	combination of MT evaluation measures as features	supervised	75.00%	82.70%
PE	Das and Smith (2009)	product of experts	supervised	76.10%	82.70%
WDDP	Wan et al. (2006)	dependency-based features	supervised	75.60%	83.00%
SHPNM	Socher et al. (2011)	recursive autoencoder with dynamic pooling	supervised	76.80%	83.60%
MTMETRICS	Madnani et al. (2012)	combination of eight machine translation metrics	supervised	77.40%	84.10%
Bi-CNN-MI	Yin and Schutze (2015)	convolutional neural network w/ multi-granular interaction	supervised	78.40%	84.60%
MP-CNN	He et al. (2015)	convolutional neural network w/ multiple perspectives	supervised	78.60%	84.73%
Recon	Cheng and Kartsaklis (2015)	Recursive NNs w/ syntax-aware multi-sense word embeddings	supervised	78.60%	85.30%
LEXDISCRIM	Ji and Eisenstein (2013)	combination of latent space and lexical features	supervised	80.41%	85.96%

https://www.aclweb.org/aclwiki/index.php?title=Paraphrase_Identification_(State_of_the_art)

Multi-instance Learning Paraphrase Model

Challenges of Twitter Data

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

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

Yes!

very short, lexically divergent

Techniques

- Multiple Instance Learning
- **Probabilistic Graphical Models**
- Markov Networks with Latent Variables

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

Assumption

*Mancini has been sacked by Manchester **City***

*Mancini gets the boot from Man **City***

Yes!

At-least-one Paraphrase Anchor

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

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

Assumption

*The **new** <u>Ciroc</u> **flavor** has arrived*
*<u>Ciroc</u> got a **new flavor** coming out*

Yes!

At-least-one Paraphrase Anchor

two sentences about the <u>same topic</u> are paraphrases
if and only if
they contain at least one word pair that is
a paraphrase **anchor** (in the context)

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

Assumption

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

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

Yes!

At-least-one Paraphrase Anchor

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

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

Assumption

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

No!

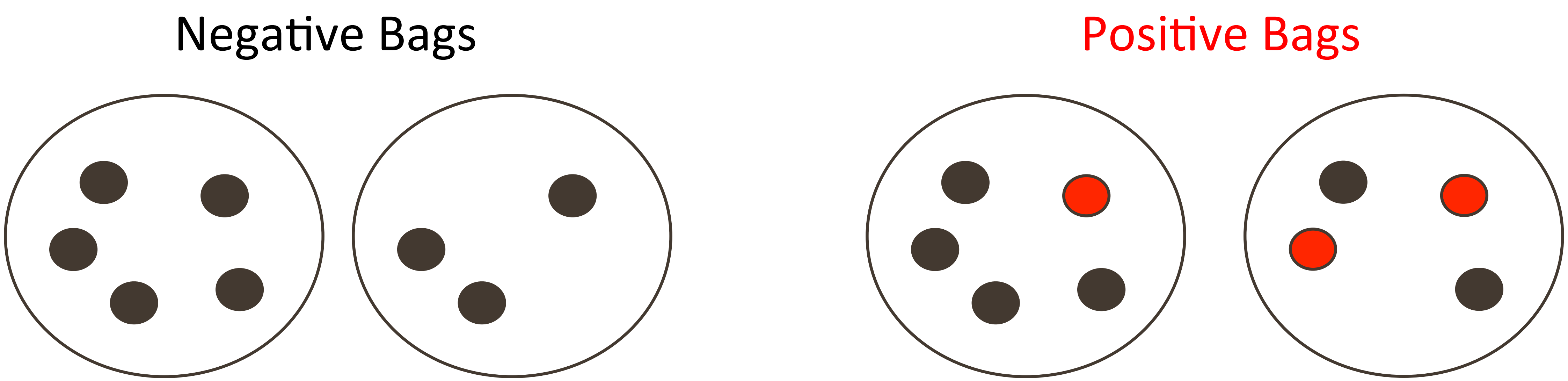
At-least-one Paraphrase Anchor

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

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

Multi-instance Learning

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

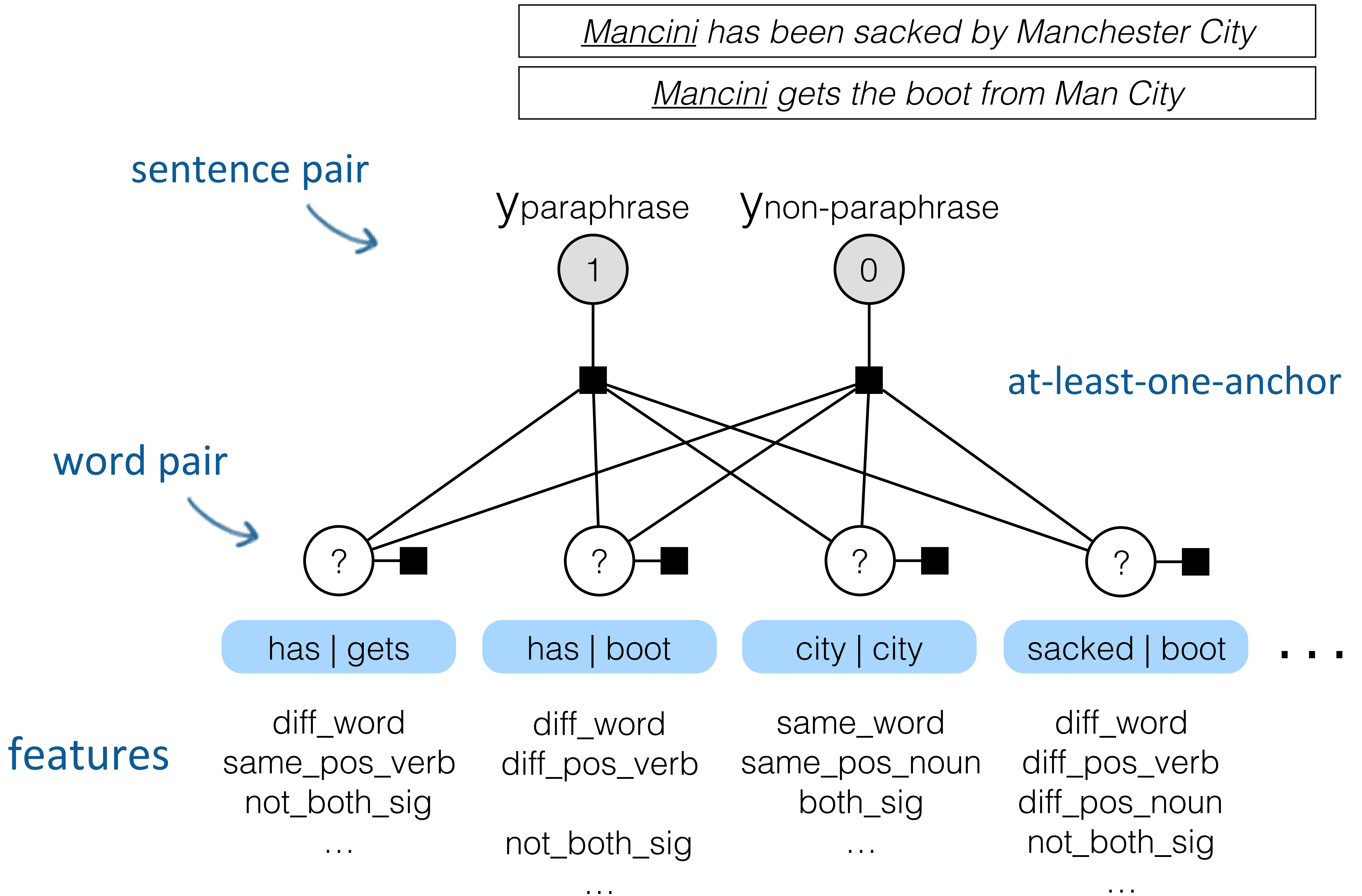


A bag is labeled negative, if
all the examples in it are negative

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

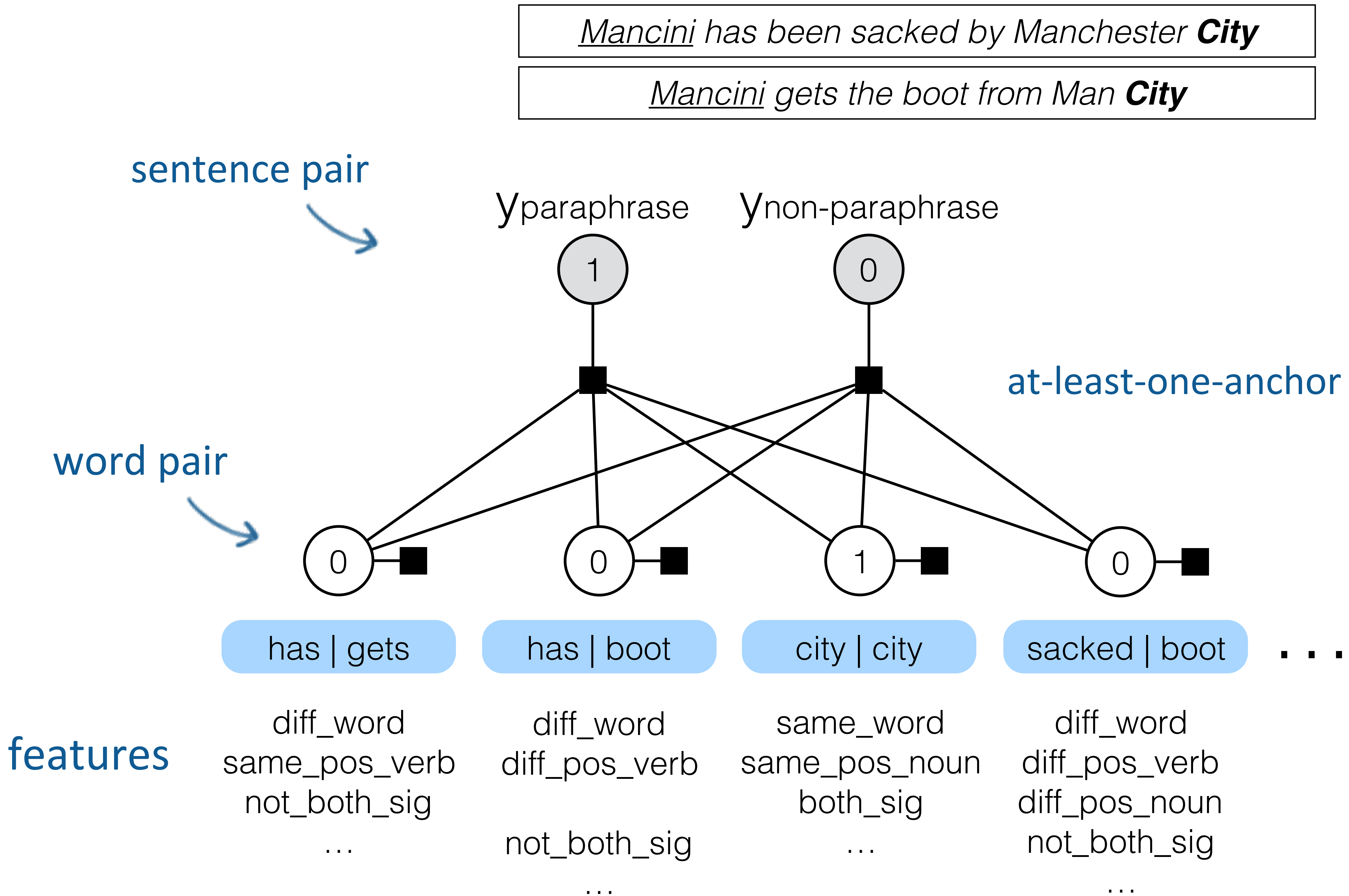
(Dietterich et al., 1997)

Multi-instance Learning Paraphrase Model



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

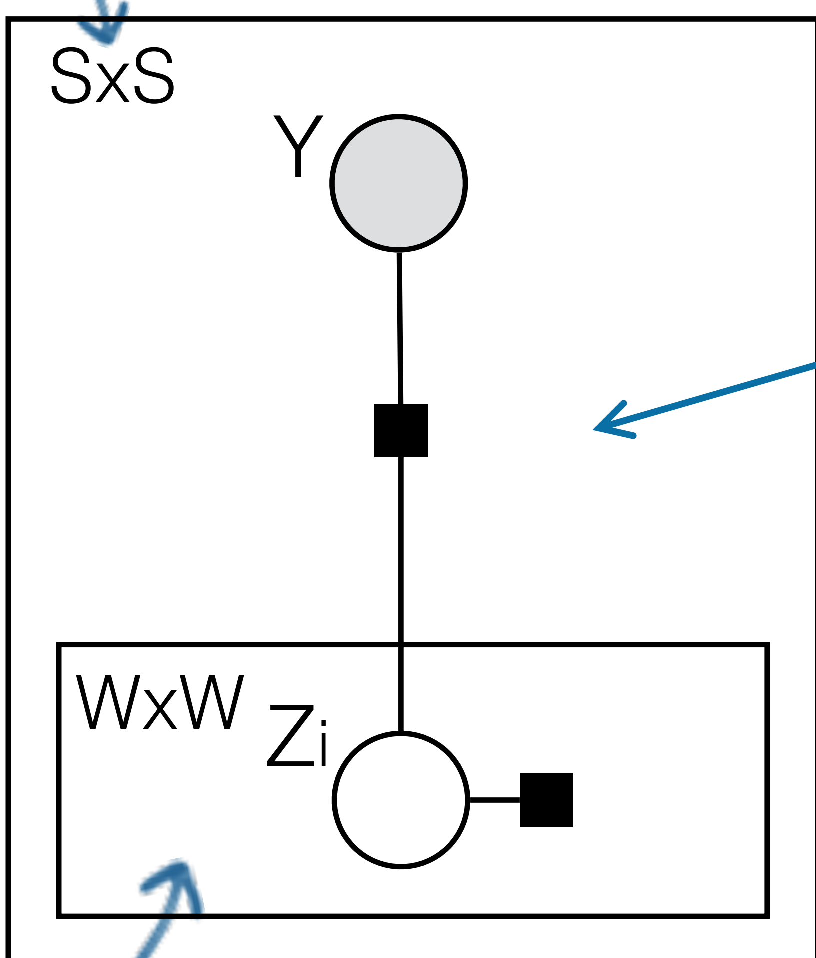


Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Multi-instance Learning

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

sentence pair



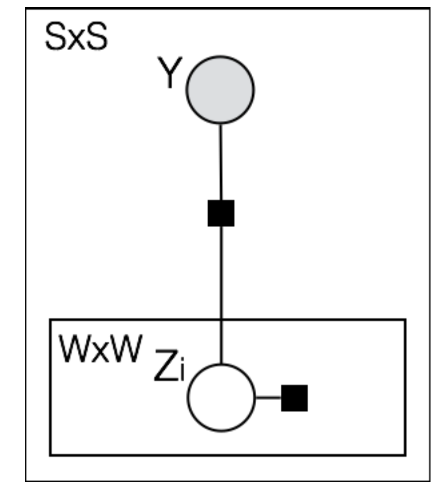
word pair

$$\sigma(\mathbf{z}, \mathbf{y}) = \begin{cases} 1 & \text{if } \mathbf{y} = \mathit{true} \wedge \exists i : z_i = 1 \\ 1 & \text{if } \mathbf{y} = \mathit{false} \wedge \forall i : z_i = 0 \\ 0 & \text{otherwise} \end{cases}$$

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

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 conditional likelihood over the training corpus

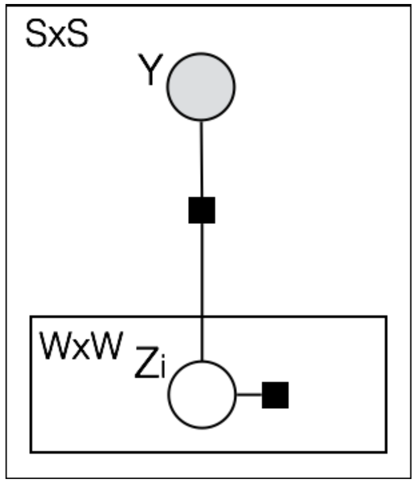
$$\theta^* = \arg\max_{\theta} P(\mathbf{y}|\mathbf{w};\theta) = \arg\max_{\theta} \prod_{i} \sum_{\mathbf{z}_i} P(\mathbf{z}_i, y_i|\mathbf{w}_i;\theta)$$

***i*th training sentence pair**

all possible values of the latent variables

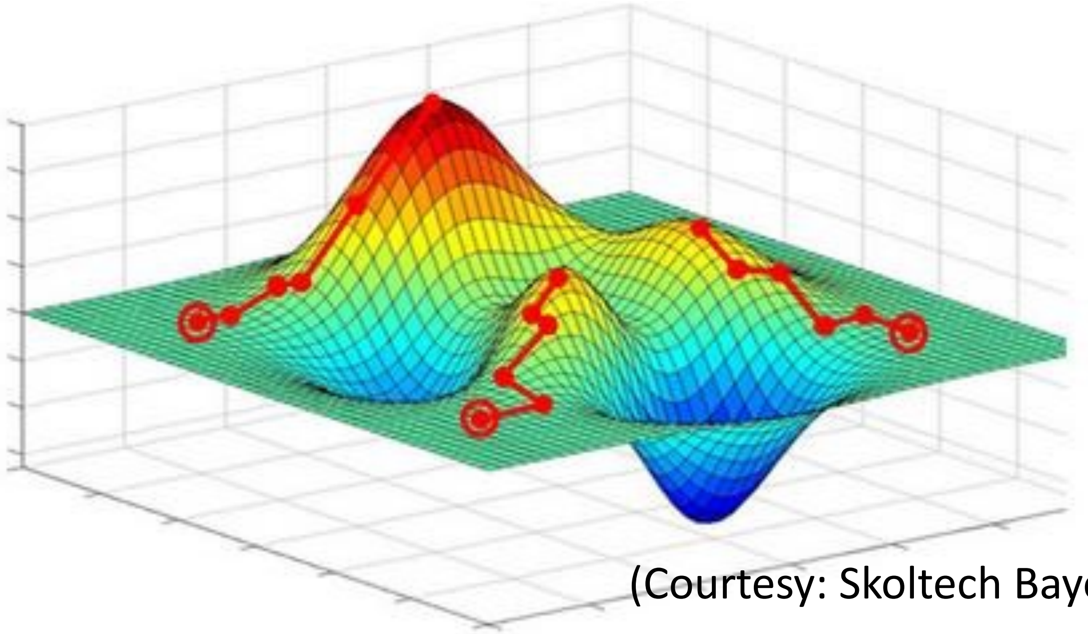
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Learning Algorithm



Finding maximum likelihood estimate:
stochastic gradient ascent

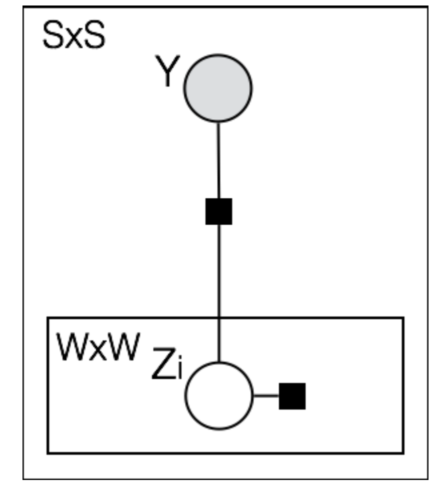
$$\frac{\partial \log P(\mathbf{y}|\mathbf{w};\theta)}{\partial\theta} = \mathbf{E}_{P(\mathbf{z}|\mathbf{w},\mathbf{y};\theta)}(\sum_i f(\mathbf{z}_i, \mathbf{w}_i)) - \mathbf{E}_{P(\mathbf{z},\mathbf{y}|\mathbf{w};\theta)}(\sum_i f(\mathbf{z}_i, \mathbf{w}_i))$$



(Courtesy: Skoltech Bayesgroup)

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)

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

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

reward correct (conditioned on labels)

penalize wrong (ignoring labels)

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

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

Twitter Paraphrase Dataset

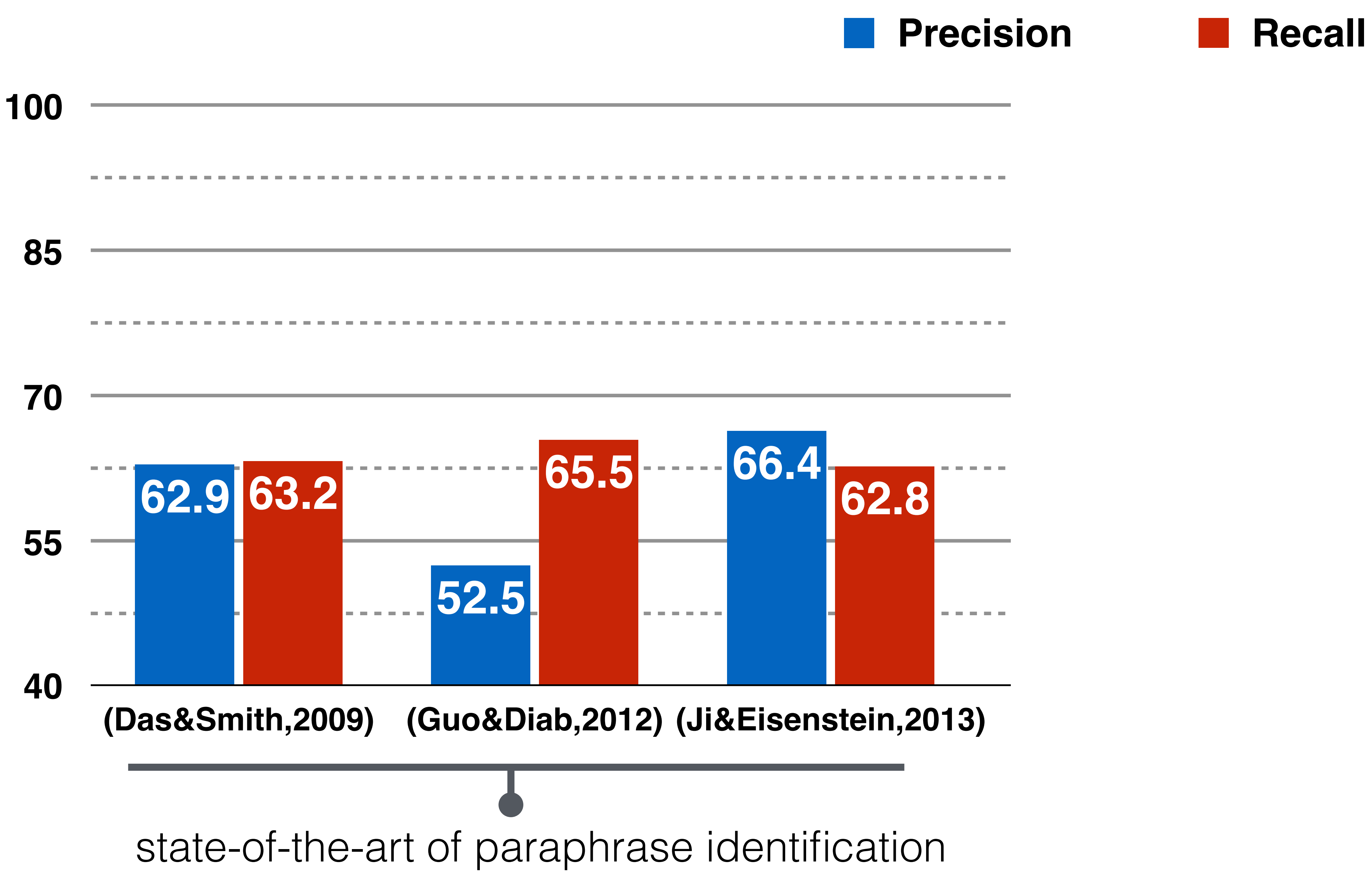
18,762 labeled sentence pairs
1/3 paraphrase, 2/3 non-paraphrase

Techniques

- **Crowdsourcing** (Human-Computer Interaction)
- SumBasic algorithm for sentence filtering
- Multi-armed bandits algorithm for topic selection

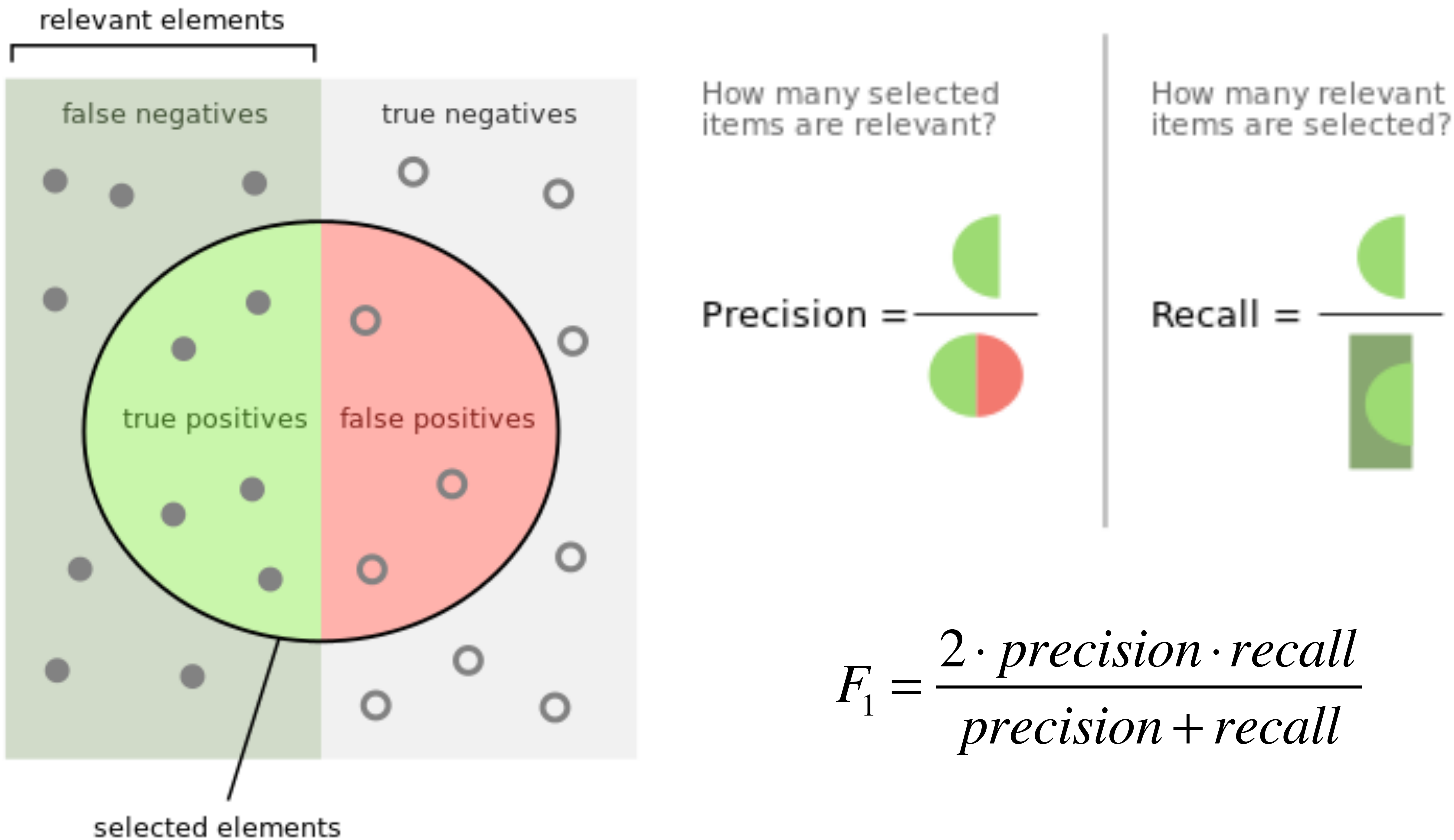
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Performance



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

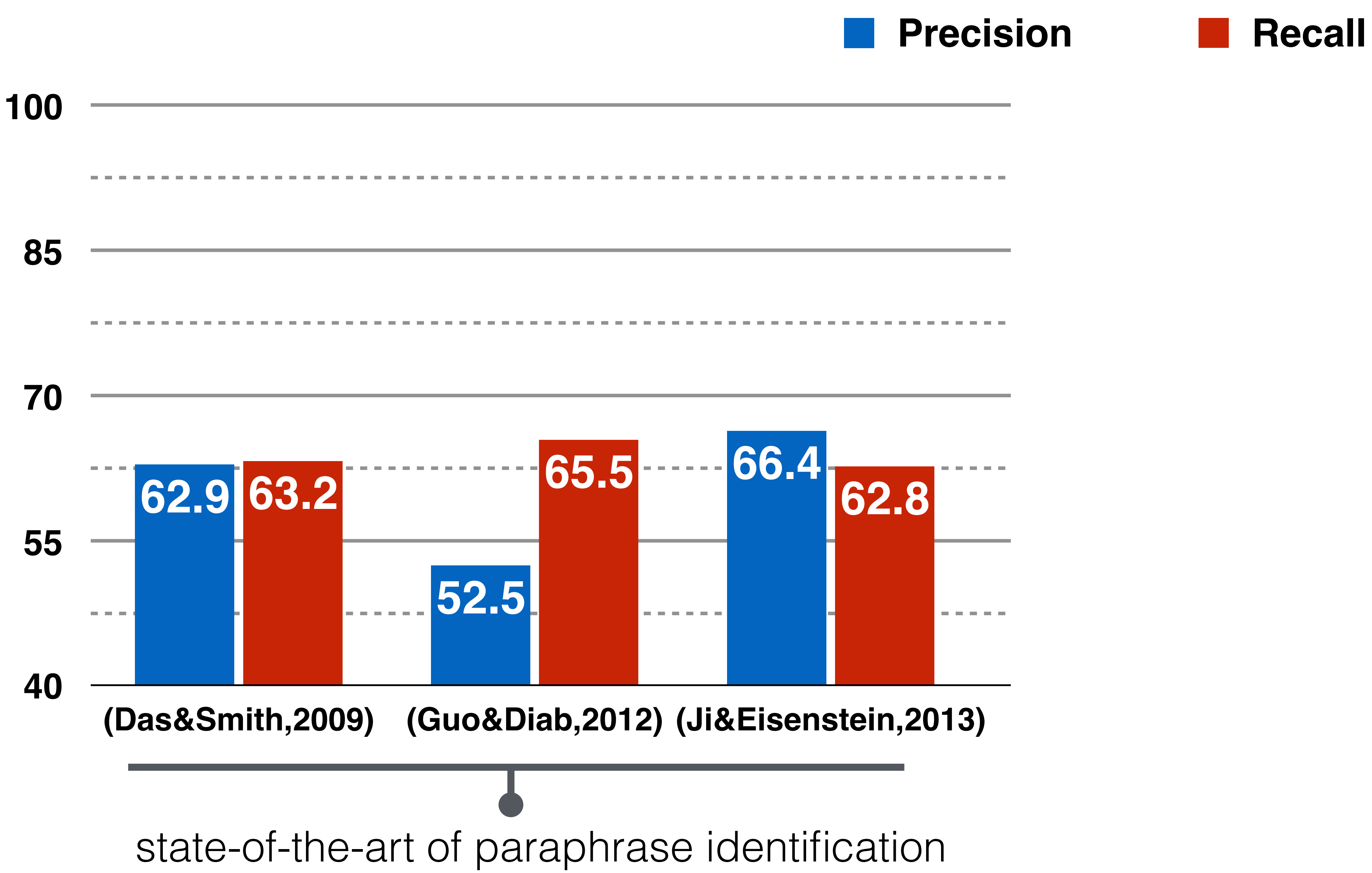
F-measure



$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

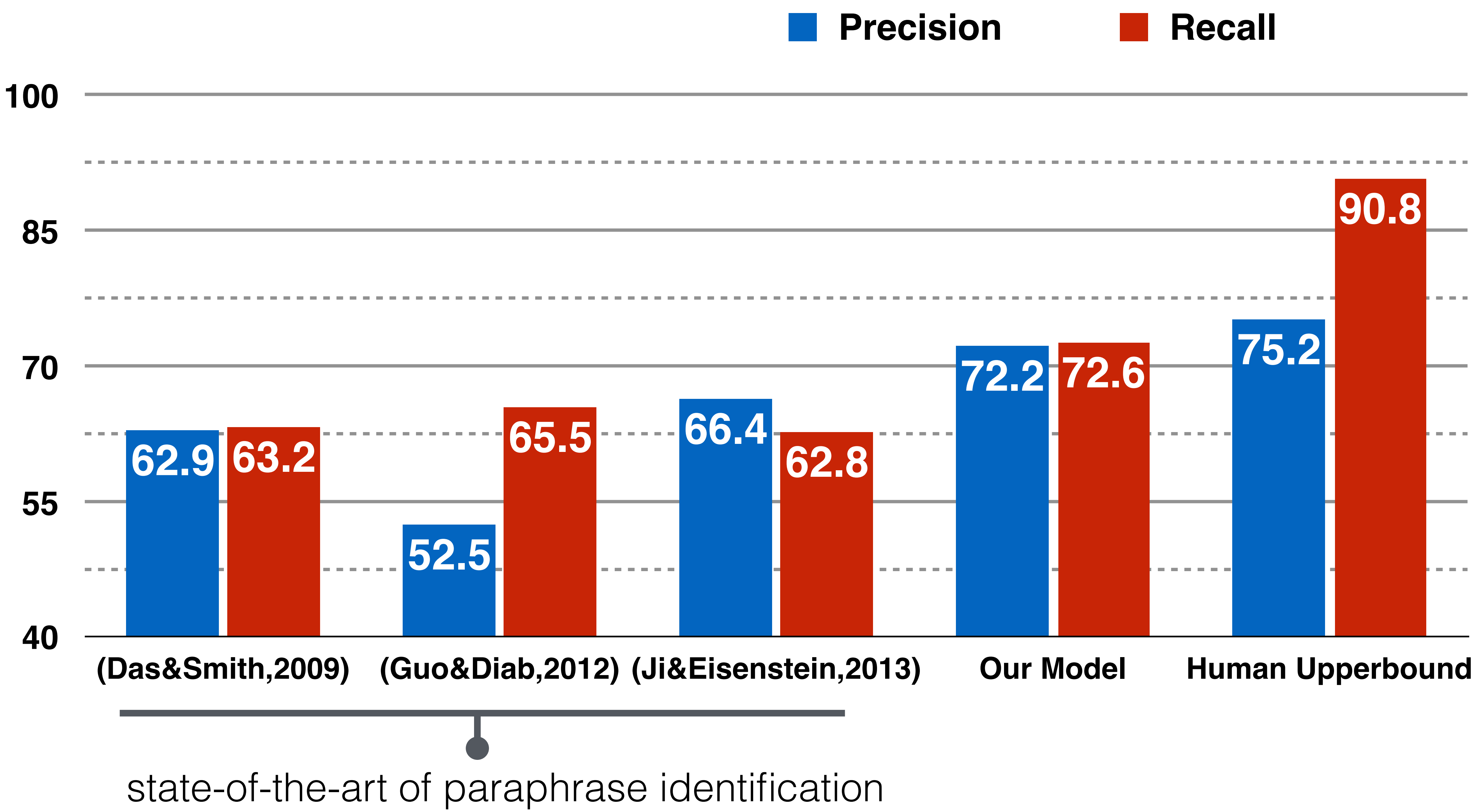
Source: Wikipedia

Performance



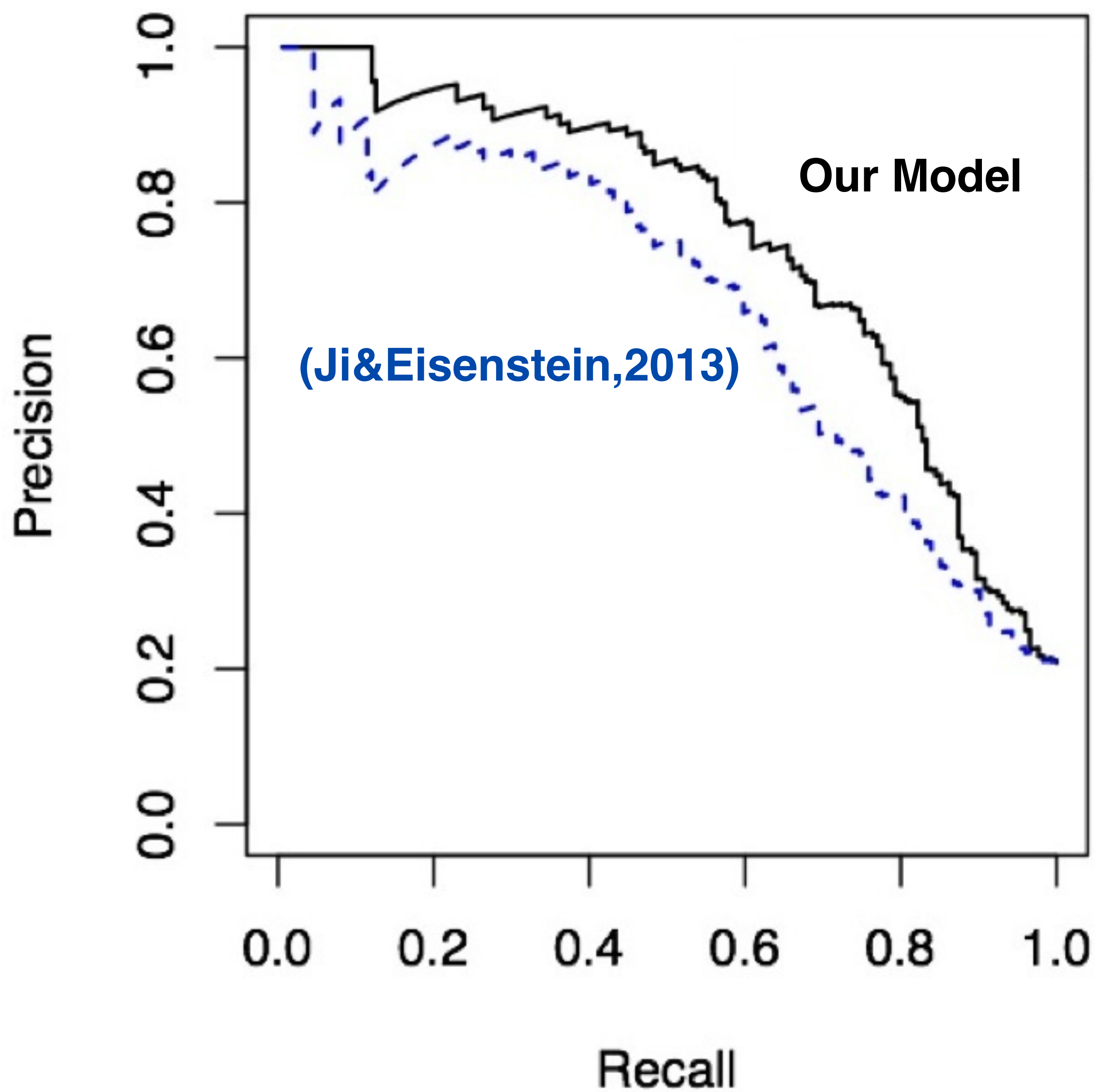
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

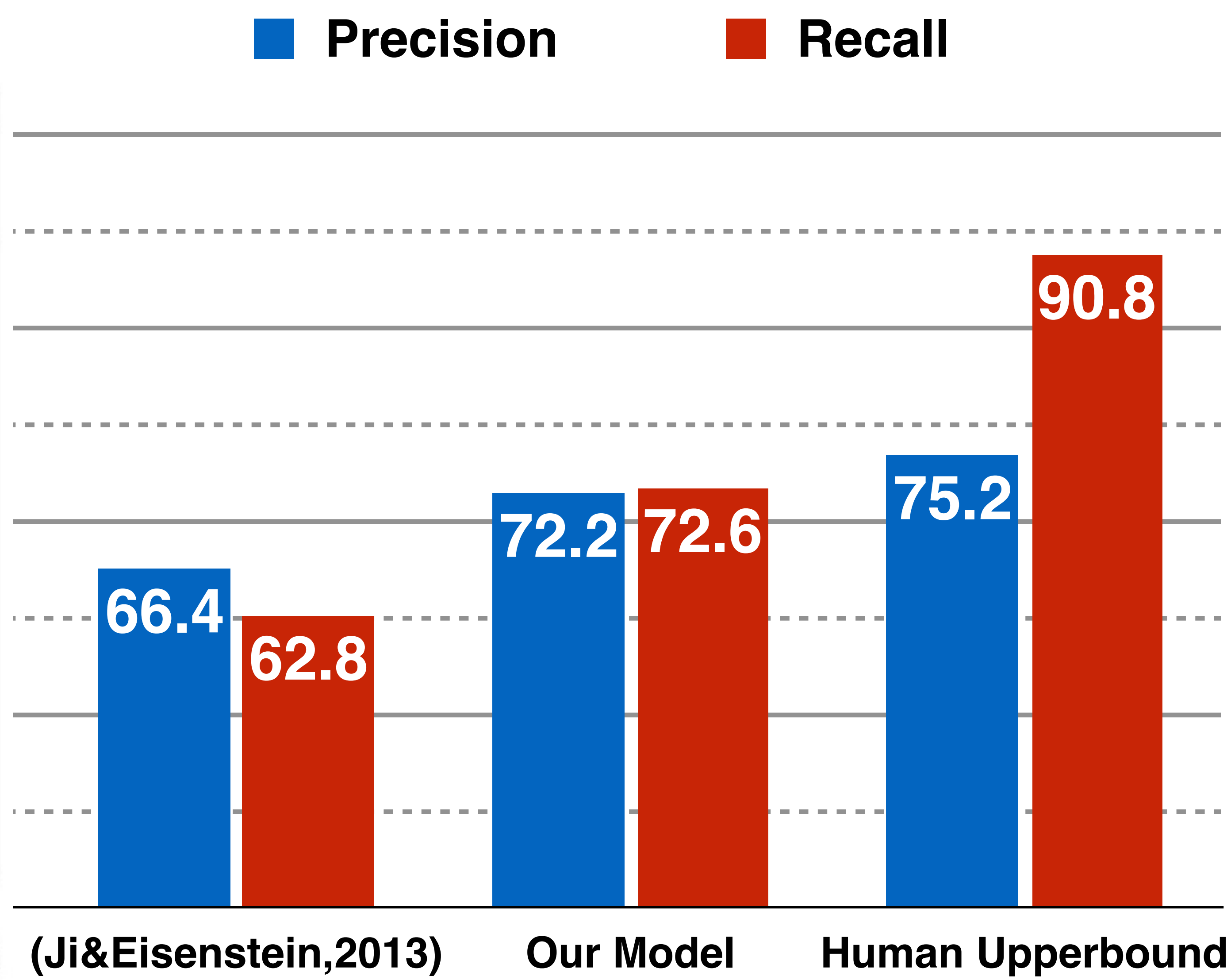
Performance



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

Performance



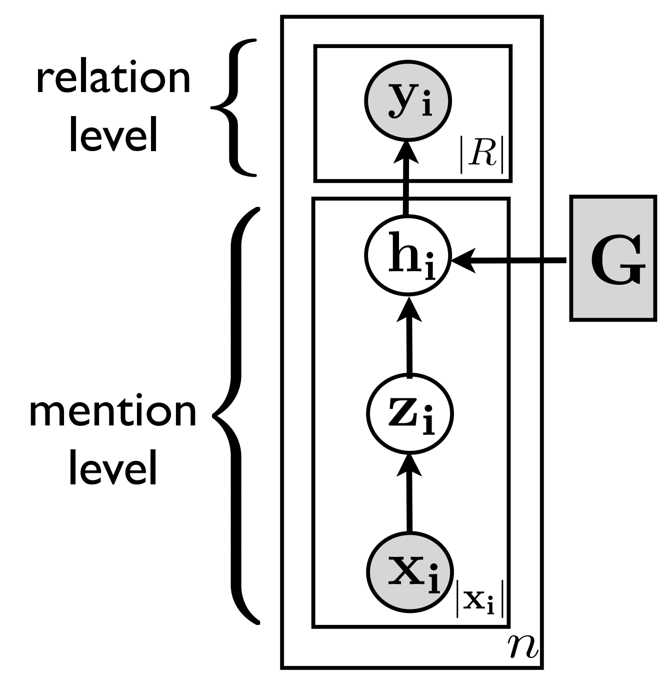


Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. “Extracting Lexically Divergent Paraphrases from Twitter” In TACL (2014)

Multi-instance Learning

Other Application: Learning Knowledge Base

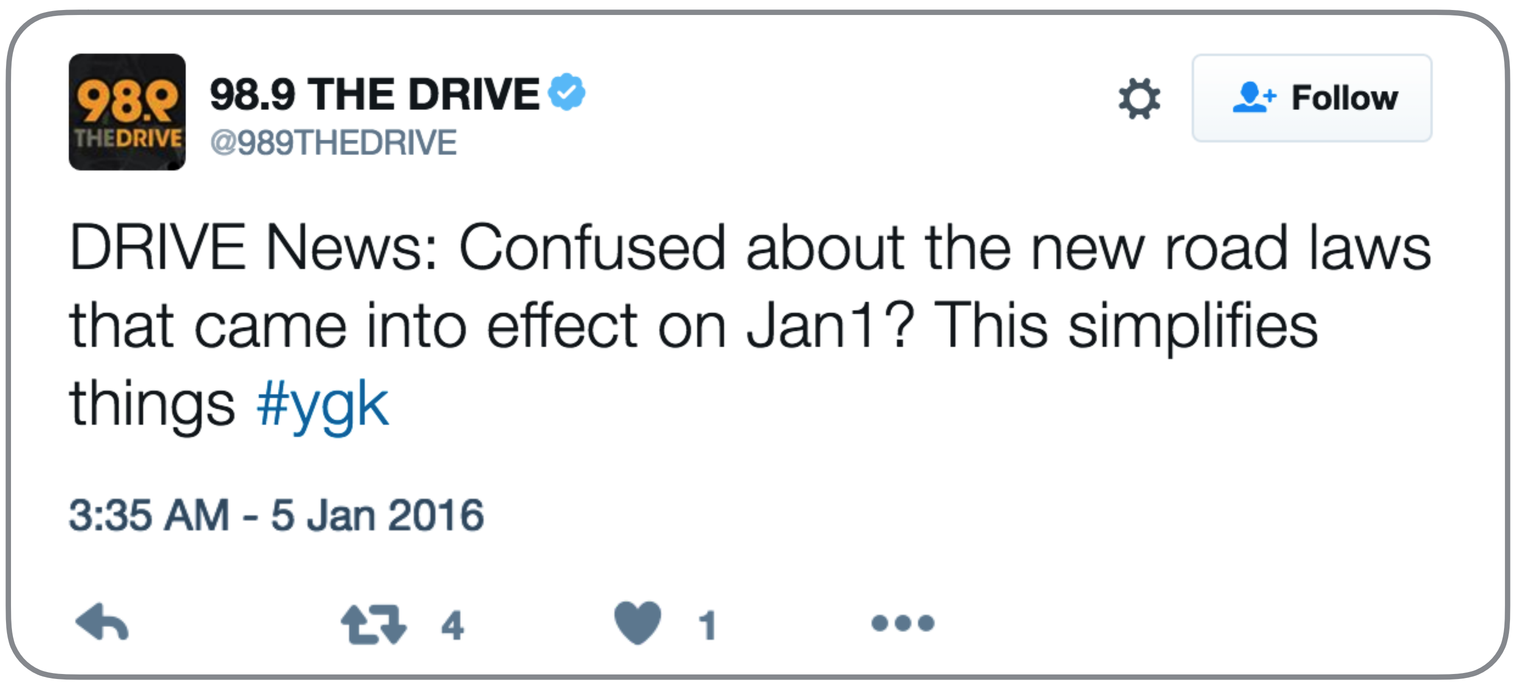
… the forced *<u>resignation</u>* *of the CEO of Boeing, Harry Stonecipher, for …*



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)

Multi-instance Learning

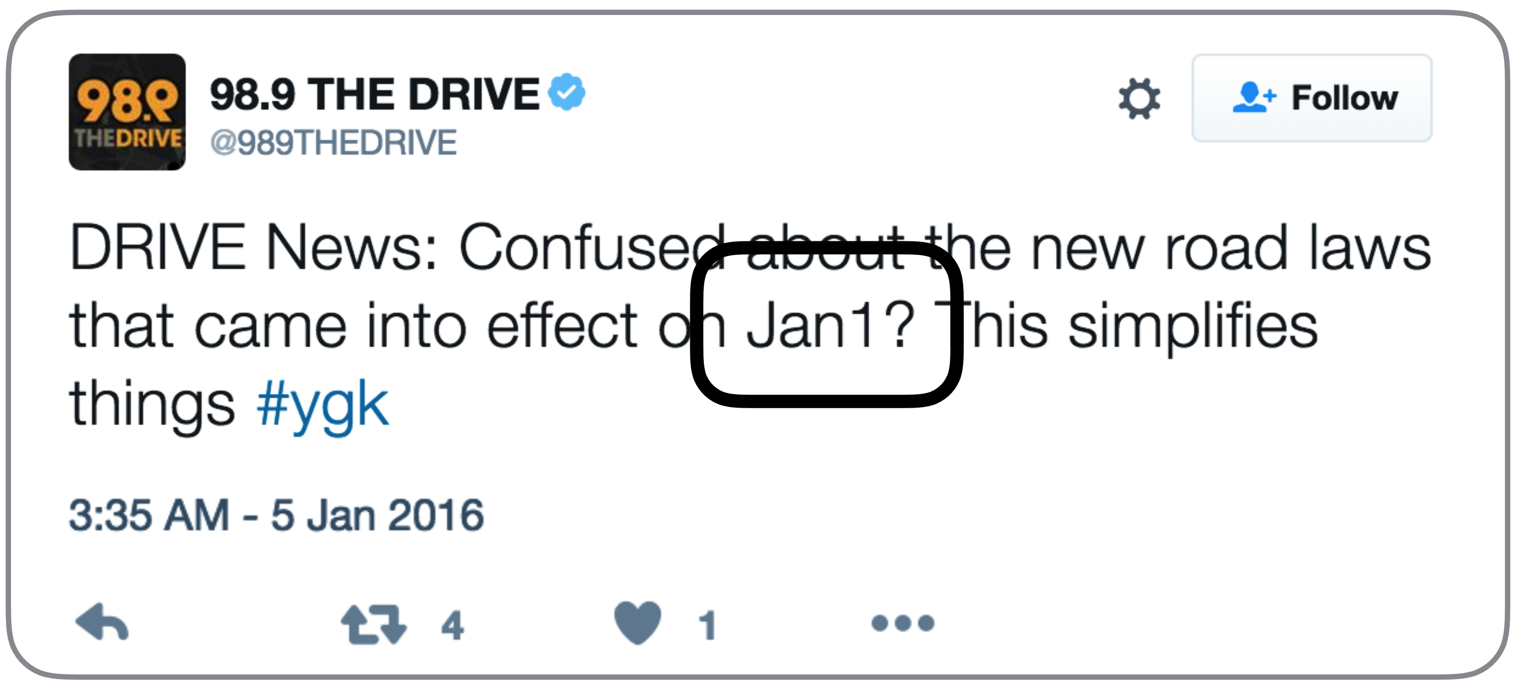
Other Application: Resolving Time Expressions



Jeniya Tabassum, Alan Ritter, Wei Xu. "A Minimally Supervised Method for Recognizing and Normalizing Time Expressions in Twitter" In EMNLP (2016)

Multi-instance Learning

Other Application: Resolving Time Expressions



Jeniya Tabassum, Alan Ritter, Wei Xu. "A Minimally Supervised Method for Recognizing and Normalizing Time Expressions in Twitter" In EMNLP (2016)

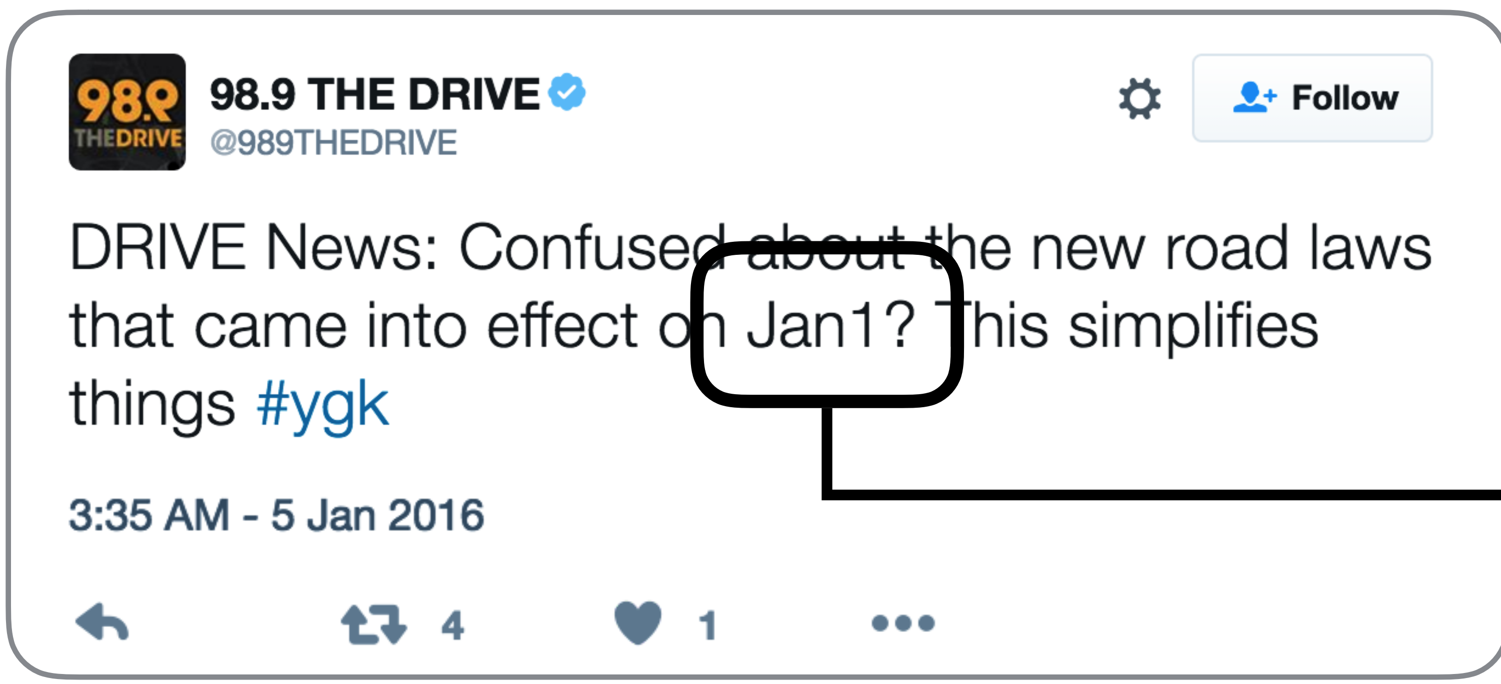
Multi-instance Learning

Other Application: Resolving Time Expressions

Jeniya Tabassum, Alan Ritter, Wei Xu. “A Minimally Supervised Method for Recognizing and Normalizing Time Expressions in Twitter” In EMNLP (2016)

Multi-instance Learning

Other Application: Resolving Time Expressions



TweeTime

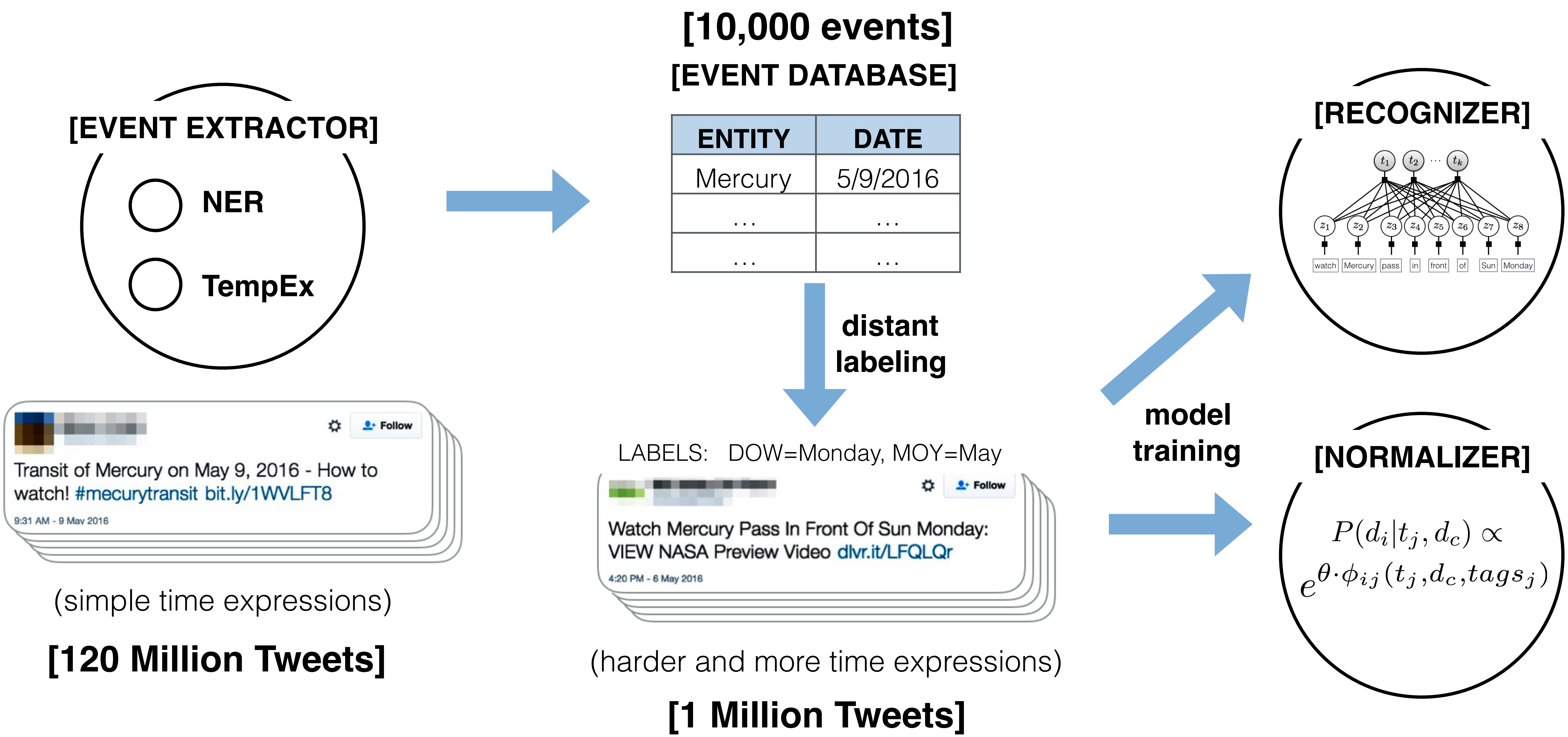
1 Jan 2016

All other state-of-the-art time resolvers { TempEX, HeidelTime, SUTime, UWTime } failed.

Jeniya Tabassum, Alan Ritter, Wei Xu. "A Minimally Supervised Method for Recognizing and Normalizing Time Expressions in Twitter" In EMNLP (2016)

Multi-instance Learning

Other Application: Resolving Time Expressions



Jeniya Tabassum, Alan Ritter, Wei Xu. “A Minimally Supervised Method for Recognizing and Normalizing Time Expressions in Twitter” In EMNLP (2016)

More about **TweeTime**

- **Jeniya Tabassum (OSU)**
- **Tuesday**, Oct 25, 12:45 pm, McPherson 2019
- A Minimally Supervised Method for Recognizing and Normalizing Time Expressions in Twitter

Another Talk Next Tuesday

- **Meg Mitchell**
- **Tuesday**, Oct 18, 3:00 pm, Dreese 480
- From Naming Concrete Objects to Sharing Abstract Thought: Vision-to-Language Begins to Grow Up

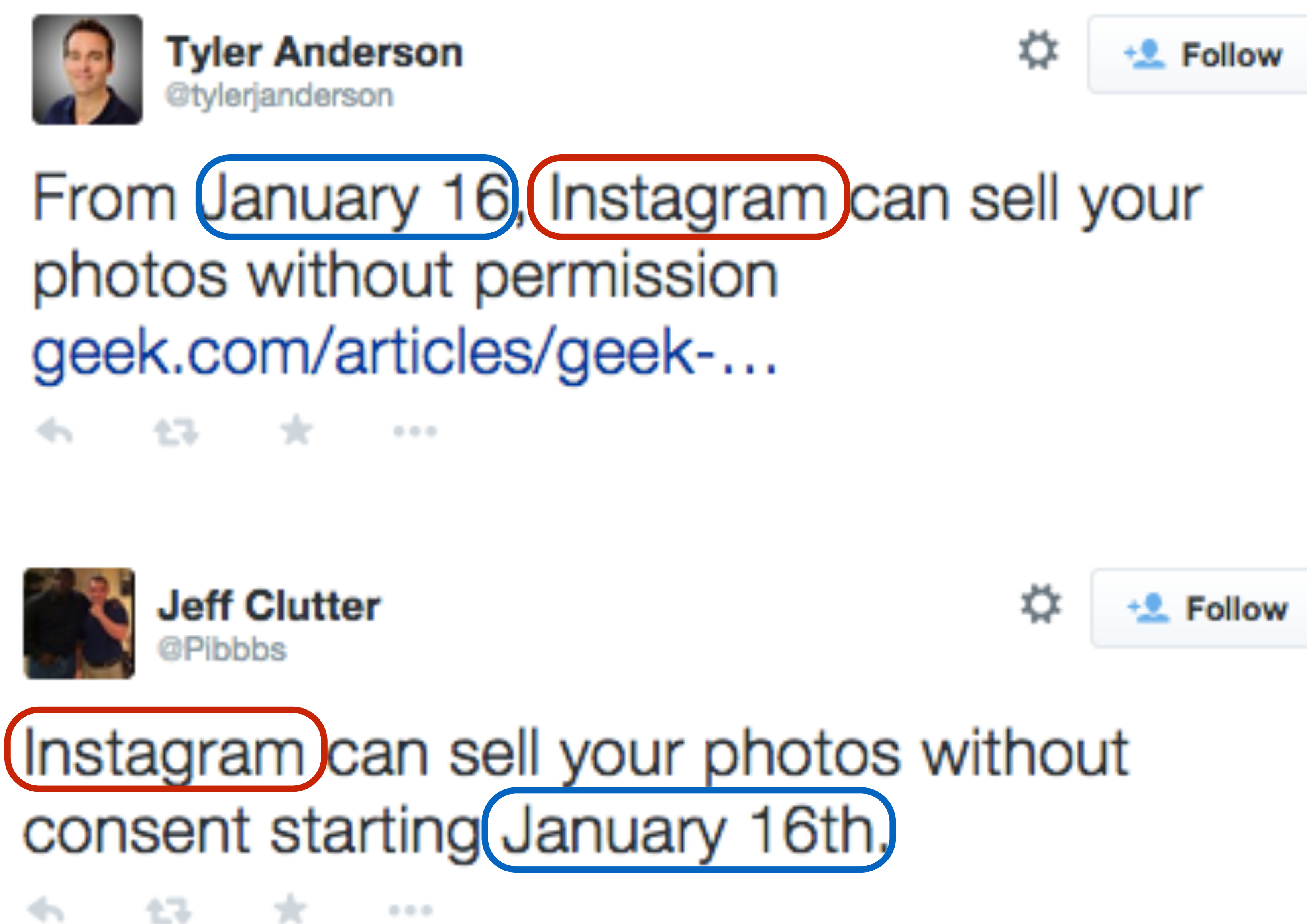
Crowdsourcing Training Data

Early Attempts on Twitter Paraphrase

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

Early Attempt:

Named Entity + Time



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

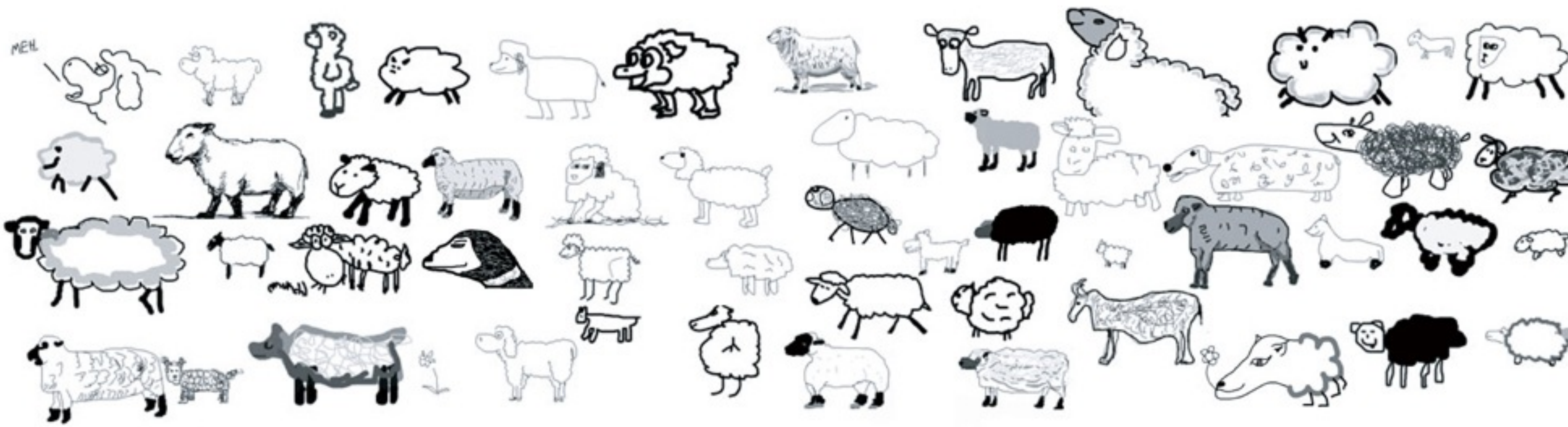
Early Attempt:

Self-translation

(Ling, Dyer, Black and Trancoso, 2013)

Crowdsourcing

art



(Courtesy: The Sheep Market by Aaron Koblin)

Crowdsourcing

paraphrase

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

Mingkun Gao, Wei Xu, Chris Callison-Burch. "Cost Optimization for Crowdsourcing Translation" In TACL (2014)

A Problem

only **8%** sentence pairs
about the same Twitter's trending topic
have similar meaning

hurts both quantity and quality

non-experts lower their bars

amazon mechanical turk™
Artificial Artificial Intelligence

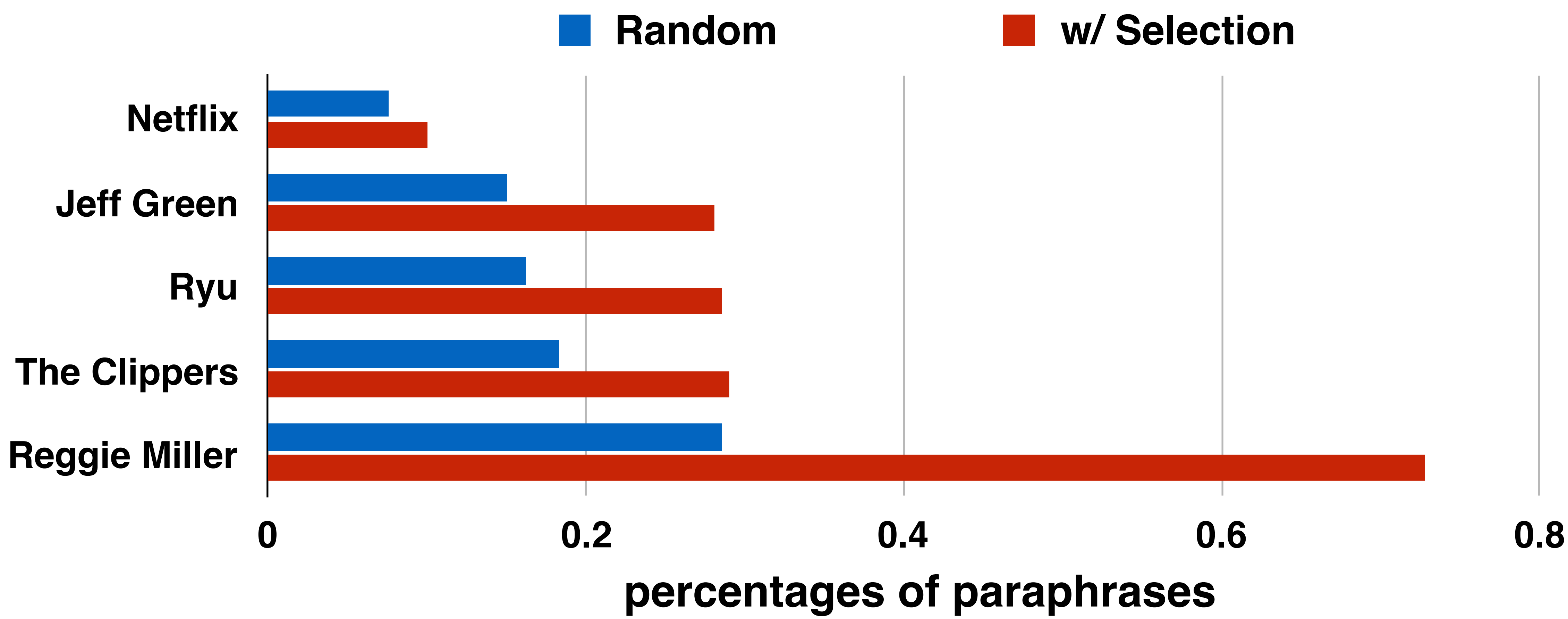
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Sentence Selection

SumBasic Algorithm

8% → 16%

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

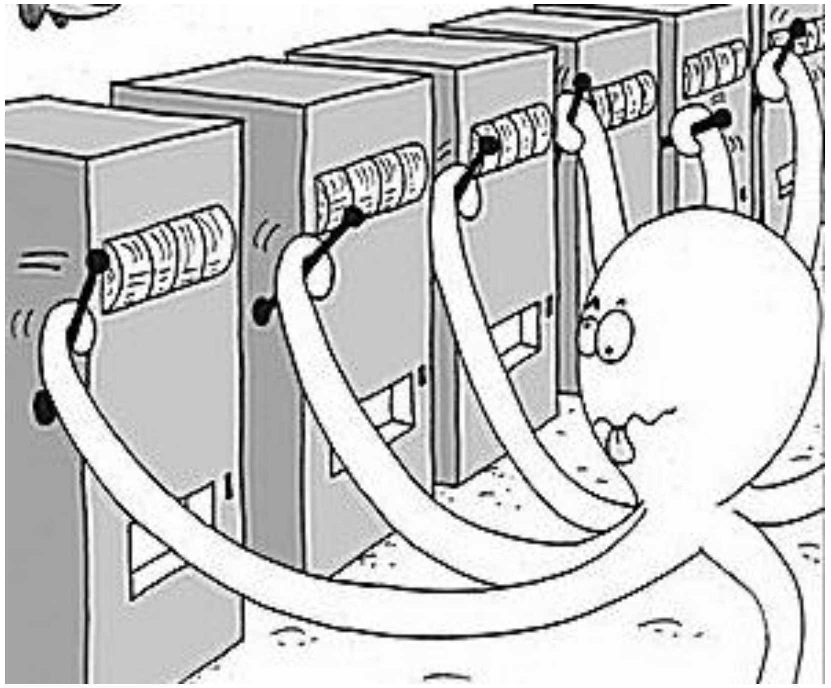


Wei Xu, Alan Ritter, Ralph Grishman. "A Preliminary Study of Tweet Summarization using Information Extraction" in LASM (2013)
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Topic Selection

Multi-Armed Bandits

16% ⟶ 34%



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

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

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

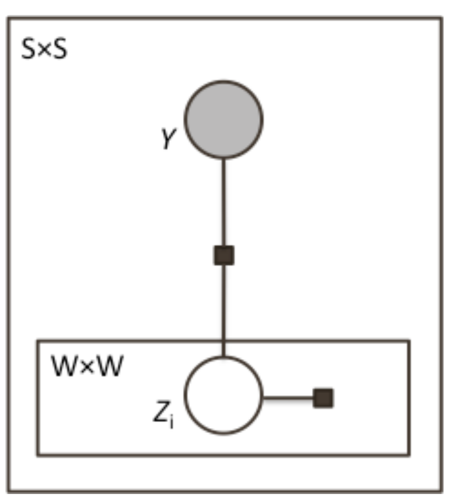
Innovations

Web-scale Paraphrase from Twitter

<u>Mancini</u> has been sacked by Manchester City
<u>Mancini</u> gets the boot from Man City

Yes!

Multi-instance Learning Paraphrase Model



- Twitter's big data stream
- joint sentence-word alignment
- no word-level annotation needed
- extensible latent variable model

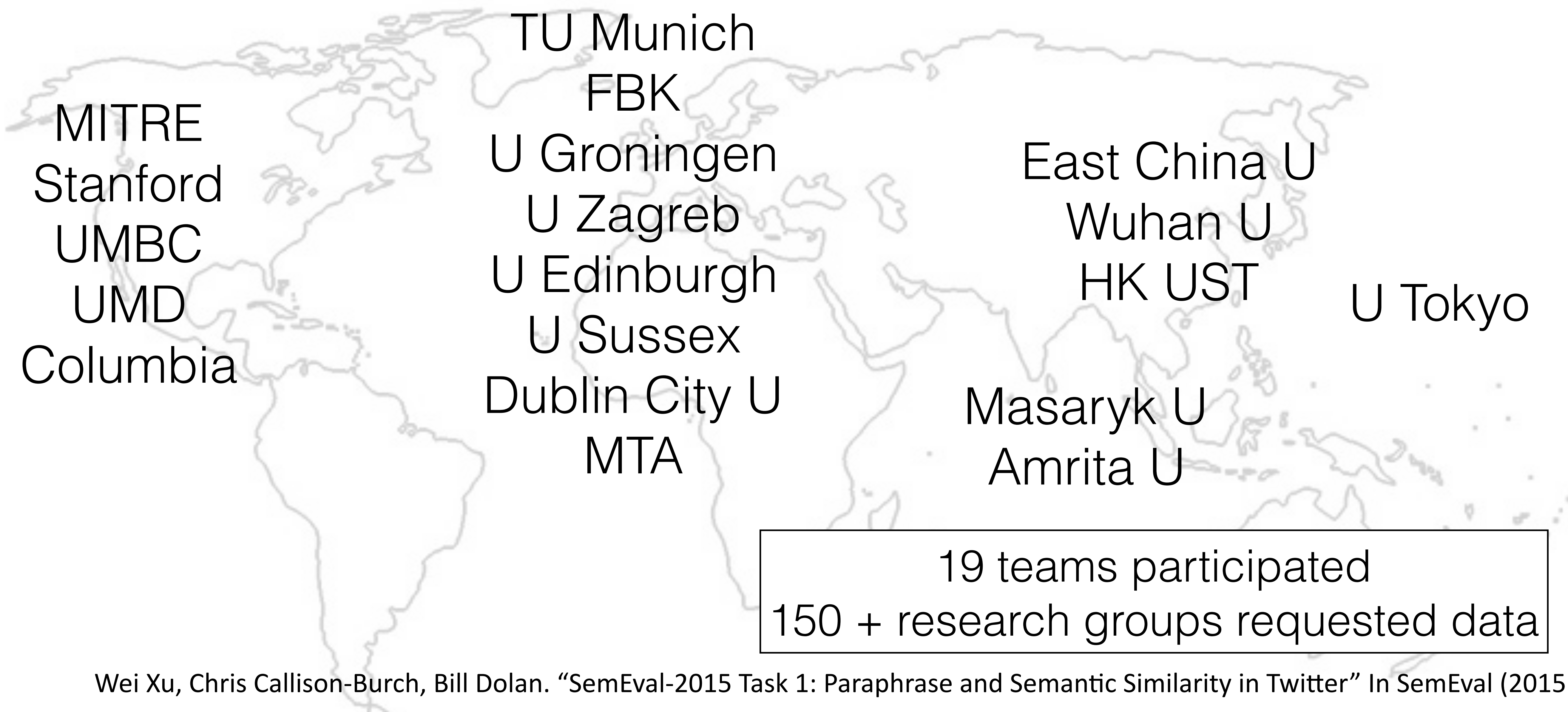
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" In TACL (2014)

Impact & Future Work

Impact

SemEval Shared Task

Paraphrase and Semantic Similarity in Twitter



19 teams participated
150 + research groups requested data

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

Challenging Cases

(Mariano "Mo" Rivera is a baseball pitcher)

Classy gesture by the <u>Mets</u> for Mariano
real class shown by the <u>Mets</u> Mo Rivera is a legend

Yes!

The <u>world of jenks</u> is such a real show
Jenks from the <u>World of Jenks</u> is such a good person

No!

Extract Phrasal Paraphrases

Mancini has been sacked by Manchester City

align

Mancini gets the boot from Man City

Techniques

- Word/Phrase Alignment
- Probabilistic Graphical Models

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)

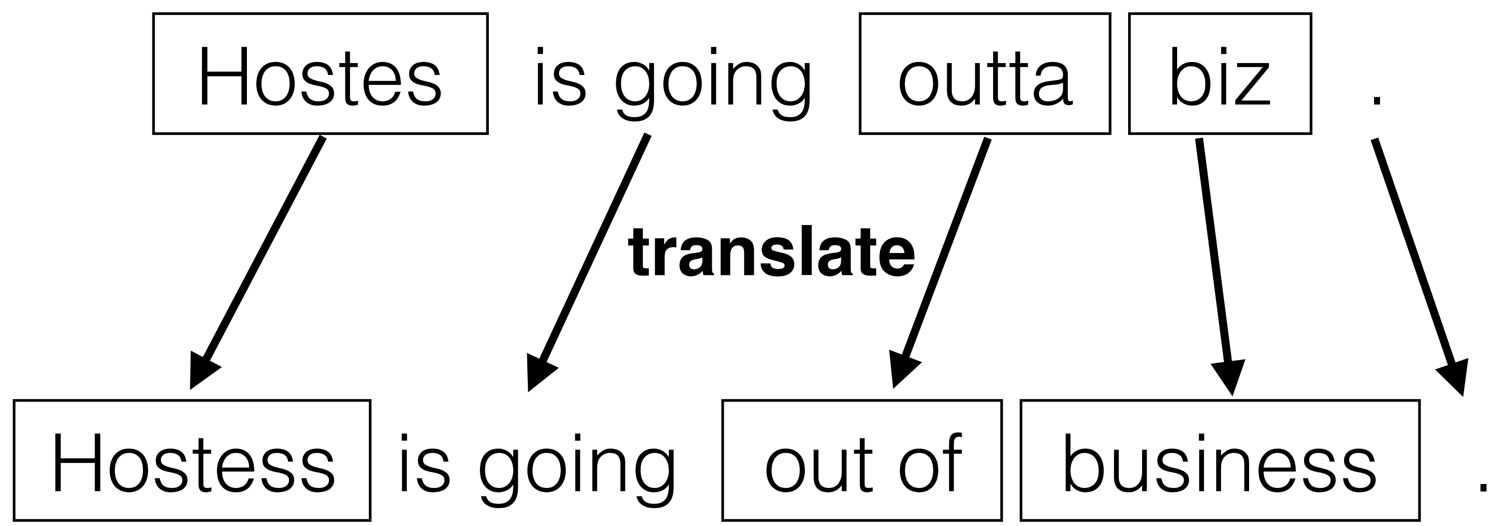
Wei Xu, Alan Ritter, Ralph Grishman. "Gathering and Generating Paraphrases from Twitter with Application to Normalization" In BUCC (2013)

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, Wei Xu. "Shared Tasks of the 2015 ACL Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition" In WNUT (2015)

Extract Phrasal Paraphrases

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

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)
Wei Xu, Alan Ritter, Ralph Grishman. "Gathering and Generating Paraphrases from Twitter with Application to Normalization" In BUCC (2013)
Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, Wei Xu. "Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition" In WNUT (2015)

Noisy Text Normalization

oscar nom'd doc

↓

Oscar-nominated documentary

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)

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

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, **Wei Xu**. "Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition" In WNUT (2015)

Natural Language Generation

	noisy	→	standard	'13* '14*
	stylistic	↔	plain	'12* '14*
	complex	→	simple	'15* '16*
	erroneous	→	correct	'11* '13*
	feminine	↔	masculine	'16*

and more (future work) …

* my research

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

Voice Assistant

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?

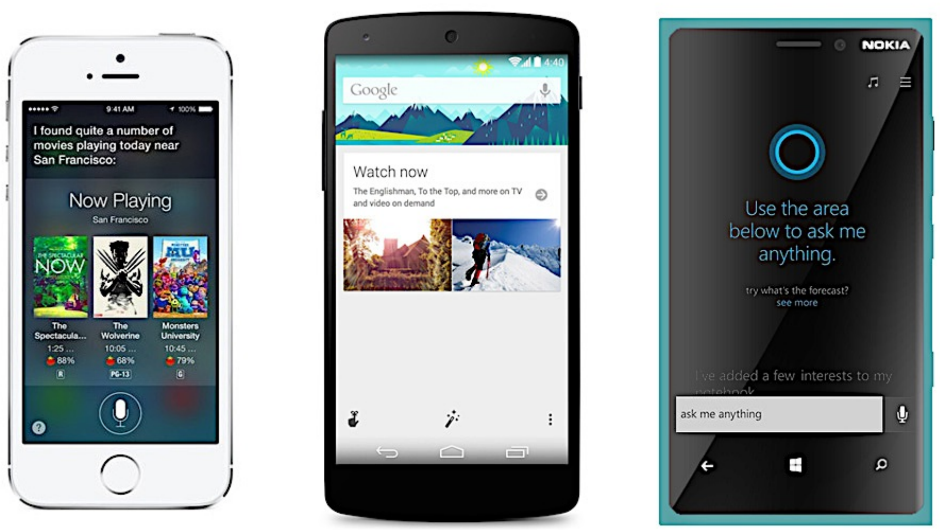
who wants to buy a beer?

who else wants to get a beer?

trying to get a beer?

… (21 different ways)

Apple Siri | Google Now | Windows Cortana



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

Unlimited Text in theory

"Almost any single (relatively complex) meaning can be implemented by an astonishingly high number of synonymous surface expressions."

Meaning-Text Linguistic Theory (Žolkovskij & Mel'čuk, 1965; ~ now)

meaning	=	invariant of paraphrases
text	=	'virtual paraphrasing'
paraphrases	=	synonymous linguistic expressions

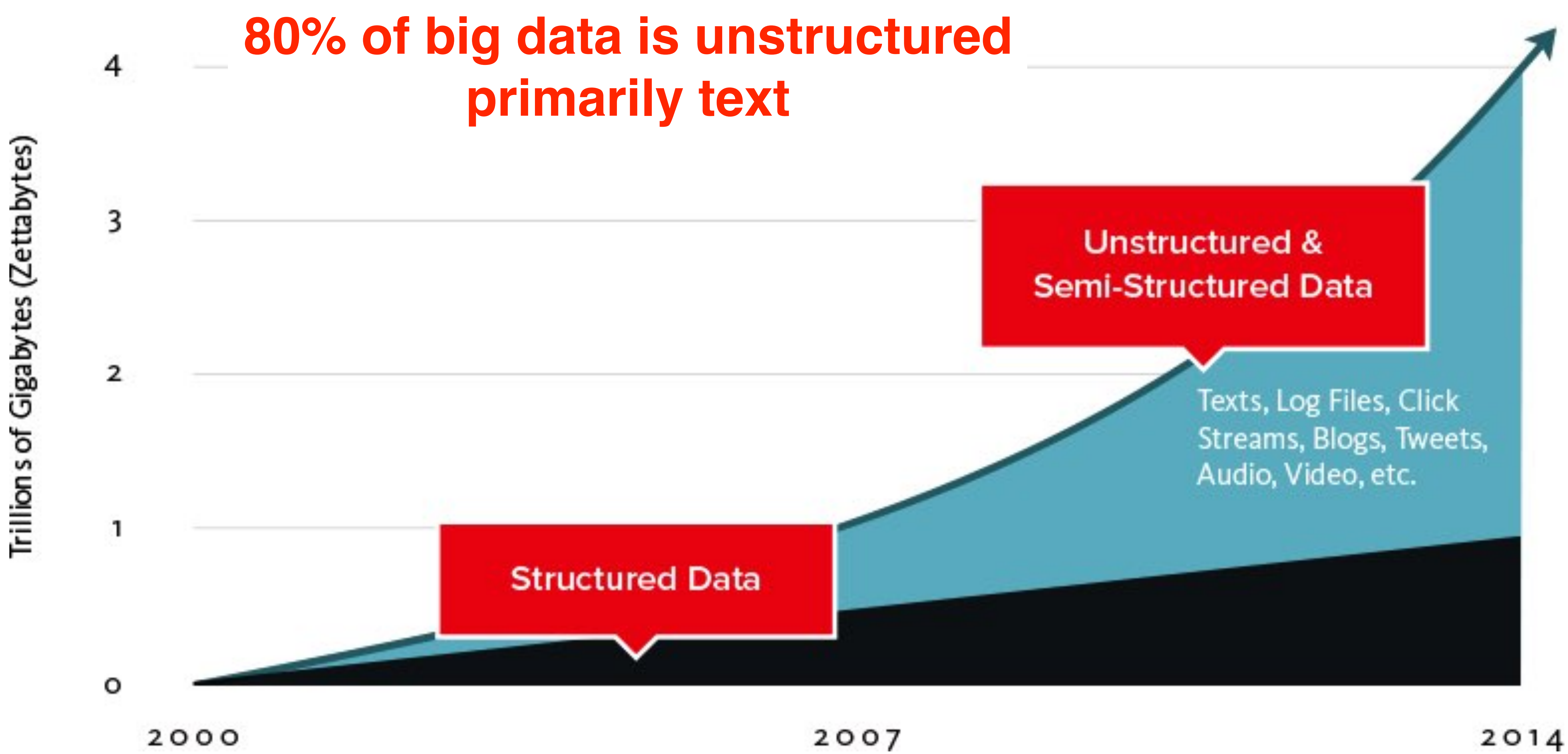
Unlimited Text in theory

“Almost any single (relatively complex) meaning can be implemented by an astonishingly high number of synonymous surface expressions.”

Meaning-Text Linguistic Theory (Žolkovskij & Mel’čuk, 1965; ~ now)

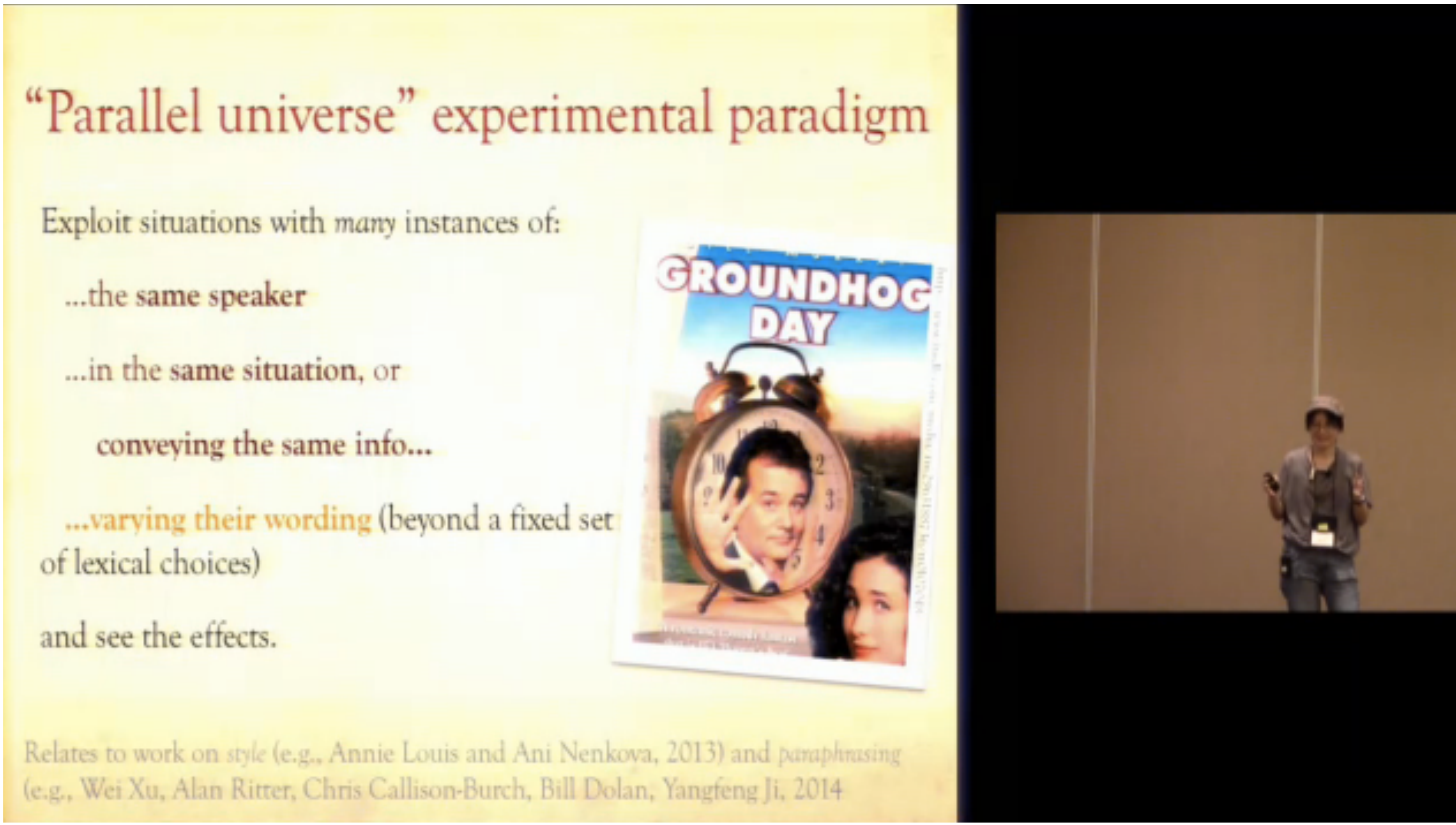
meaning	=	invariant of **paraphrases**
text	=	‘virtual **paraphrasing**’
paraphrases	=	synonymous linguistic expressions

Unlimited Text in practice



(Source: IDC Research & Couchbase)

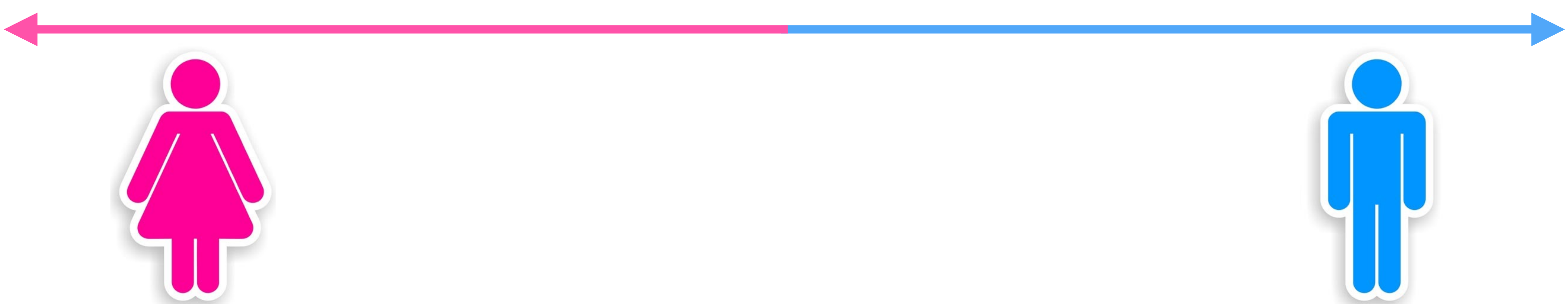
Social Science



(Lillian Lee's Keynote Speech at NAACL 2015)

Social Science

wonderfully delightfully beautifully fine well good nicely superbly



she says **he says**

(also age & income)

Daniel Preotiuc, Wei Xu, Lyle Ungar. "Discovering User Attribute Stylistic Differences via Paraphrasing" in AAAI (2016)

Social Media 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***

Language Education

Aaaaaaaaand Stephen Curry ***is on fire***

↕

What an incredible performance *from Stephen Curry*

Listen & Speak
Like a Native Speaker

thank u 4 ur time

thanku

Thank You

thanking you

appreciate it

gratitude

thx

3x

tyvm

thanks

thank you very much

say thanks

thnx

wawwww thankkkkkkkkkk you alottttttttttt!

thanks a lot

I am grateful

socialmedia-class.org