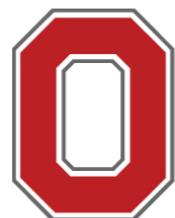


# Can paraphrase be a ultimate solution for NLU and NLG?

Wei Xu

Department of Computer Science and Engineering



THE OHIO STATE UNIVERSITY

# Can paraphrase be a ultimate solution for NLU and NLG? (or one of the solutions)

Wei Xu

Department of Computer Science and Engineering



THE OHIO STATE UNIVERSITY



We need a selfie-stick-like solution for the “selfie” problem.

a.k.a UNK problem in ever evolving human language

# The Ultimate NLP Quest: Unlimited Text

“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)

# The Ultimate NLP Quest: Unlimited Text

“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

# My take on Unlimited Text

learn and model very-large-scale paraphrases

*selfie*

**word**

*photo*

*gets the boot from*

**phrase**

*has been sacked by*

*Mr Corbyn is actually a  
secret supporter of Brexit.*

**sentence**

*Jeremy Corbyn is a closest  
Brexiteer.*

# What's good about Paraphrases ?

**fundamentally useful for a wide range of applications**

# What's good about Paraphrases ?

**fundamentally useful for a wide range of applications**

**e.g. Question Answering**

Who is the CEO stepping down from Boeing?

# What's good about Paraphrases ?

**fundamentally useful for a wide range of applications**

## e.g. 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 ...*

# Web is a Gold Mine



The New York Times @nytimes · 12 Oct 2016

Worries over the health of King Bhumibol Adulyadej are shaking Thailand  
[nyti.ms/2dRzPcr](http://nyti.ms/2dRzPcr)



5

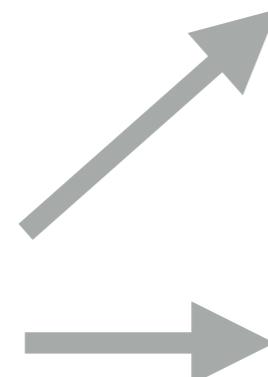
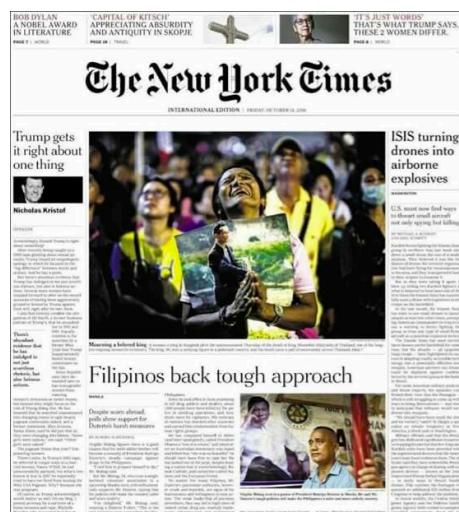


261



144

# Web is a Gold Mine



 **The New York Times**  @nytimes · 12 Oct 2016  
Worries over the health of King Bhumibol Adulyadej are shaking Thailand  
[nyti.ms/2dRzPcr](http://nyti.ms/2dRzPcr)

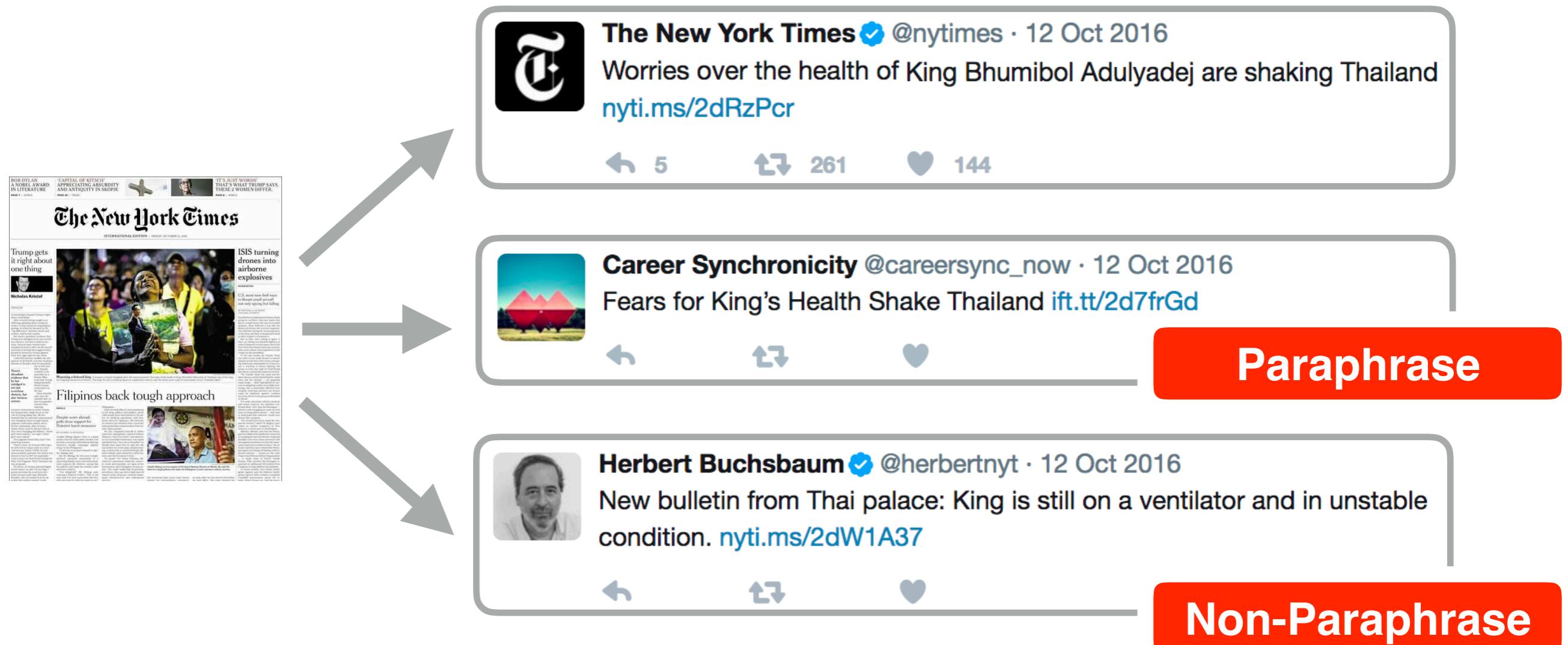
5 261 144

 **Career Synchronicity** @careersync\_now · 12 Oct 2016  
Fears for King's Health Shake Thailand [ift.tt/2d7frGd](http://ift.tt/2d7frGd)

5 261 144

**Paraphrase**

# Web is a Gold Mine



# Automatic Paraphrase Identification + Word Alignment

- Streaming data + Unsupervised model (Xu et al. 2013)
- Topic detection + Multiple Instance Learning (Xu et al. 2014)
- URL linked data + Deep Pairwise Model (Lan et al. 2017)
- Ongoing work ...

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

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

Wuwei Lan, Siyu Qiu, Hua He, Wei Xu. "A Continuously Growing Dataset of Sentential Paraphrases" in EMNLP (2017)

Wuwei Lan, Wei Xu. "A Better Pairwise Neural Model" Ongoing Work

# Automatic Paraphrase Identification + Word Alignment

- Streaming data + Unsupervised model (Xu et al. 2013)
- Topic detection + Multiple Instance Learning (Xu et al. 2014)
- URL linked data + Deep Pairwise Model (Lan et al. 2017)
- Ongoing work ...

Das & Smith 2014; Sicher et al. 2011; Ling et al. 2013; Ji & Eisenstein 2013; Parikh et al. 2016; Witting & Gimpel 2017]

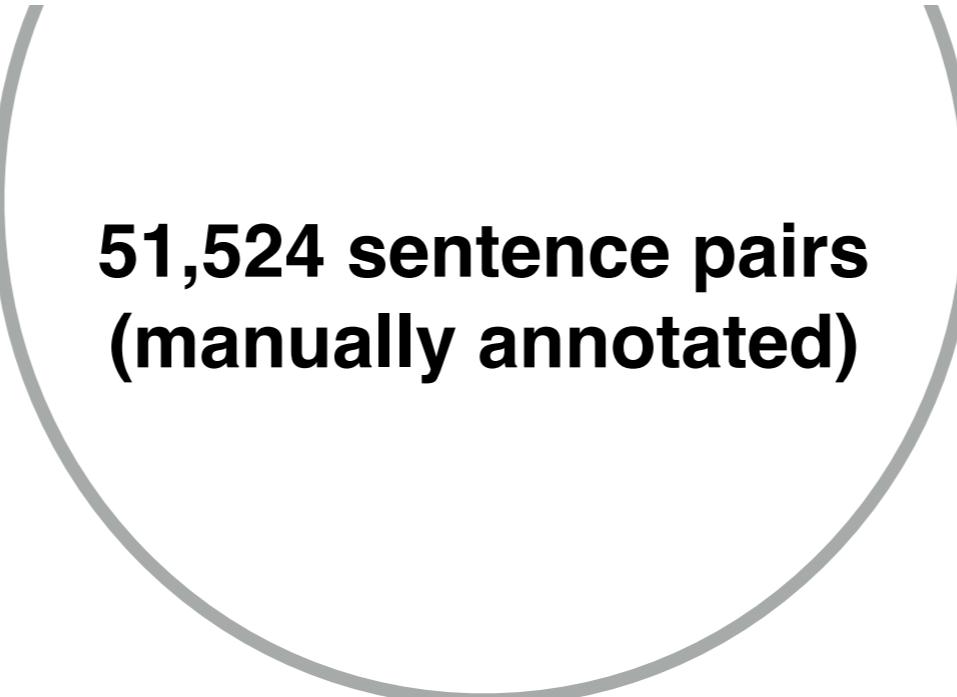
Wei Xu, Alan Ritter, Ralph Grishman. "Gathering and Generating Paraphrases from Twitter with Application to Normalization" In BUCC (2013)

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

Wuwei Lan, Siyu Qiu, Hua He, Wei Xu. "A Continuously Growing Dataset of Sentential Paraphrases" in EMNLP (2017)

Wuwei Lan, Wei Xu. "A Better Pairwise Neural Model" Ongoing Work

# [Twitter Paraphrase Corpus]



**51,524 sentence pairs  
(manually annotated)**

# [Twitter Paraphrase Corpus]

**51,524 sentence pairs  
(manually annotated)**

**> 30,000 new sentential paraphrases  
every month (automatically harvested)**

# Timely Paraphrases

Donald Trump, DJT, Drumpf, Mr Trump, Idiot Trump,  
Chump, Evil Donald, #OrangeHitler, Donald  
@realDonaldTrump, 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

# Some Cherries

*Yemeni Rebels Again Fire at US Warship in Red Sea*

*Houthi Rebels Fire 2 Missiles At Us Navy Destroyer In Persian Gulf*

*University of Virginia dean*

*UVA administrator*

*UVa official*

*U-Va. dean*

# Noisy Text Normalization

Hostes is going outta biz .



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

Hostes is going outta biz .



**translate**

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



Hostes is going outta biz .  
**translate**

A diagram illustrating the process of noisy text normalization. On the left, the Twitter logo is shown above a noisy tweet: "Hostes is going outta biz .". A large arrow points from the word "Hostes" down to the word "translate" in bold text, indicating that the normalization process involves translating the noisy text into cleaner, normalized text.

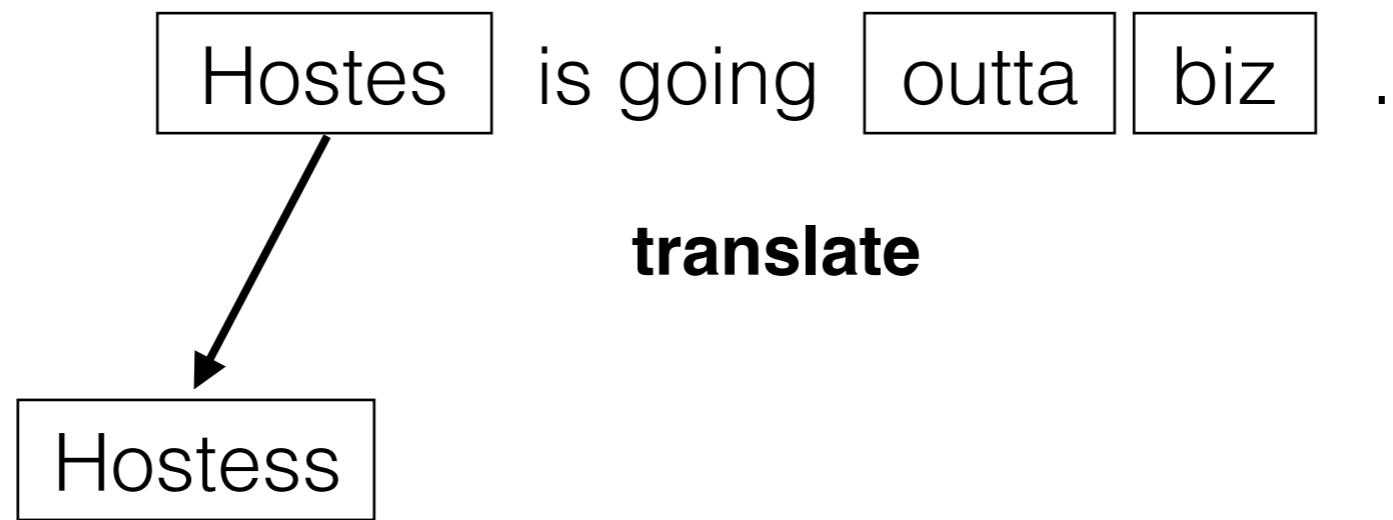
Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao.

“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



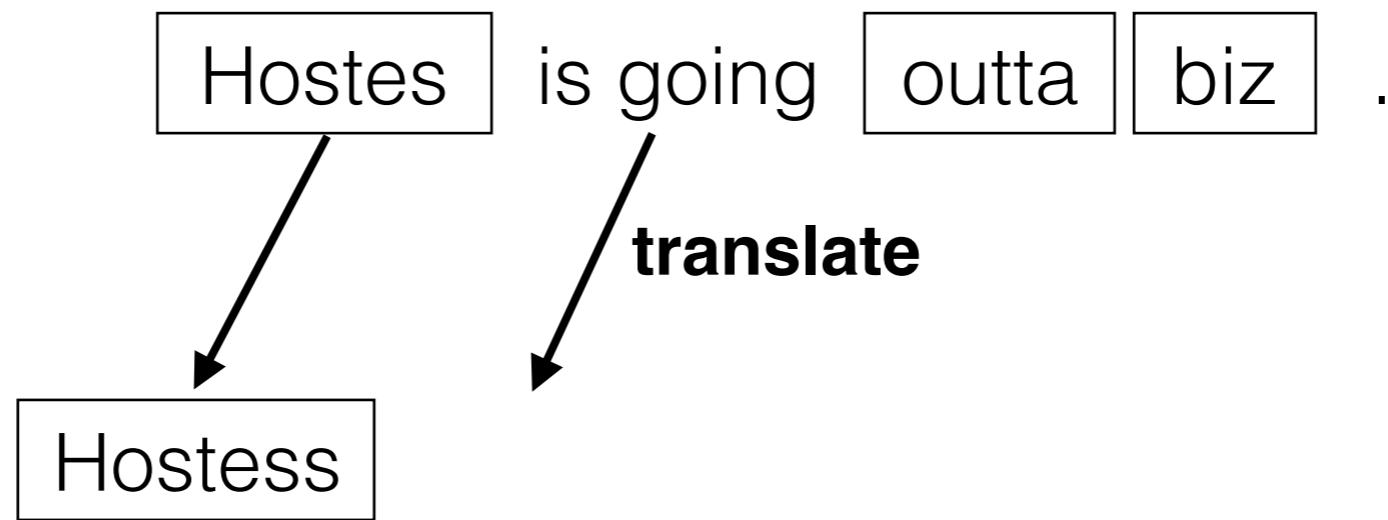
Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao.

“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



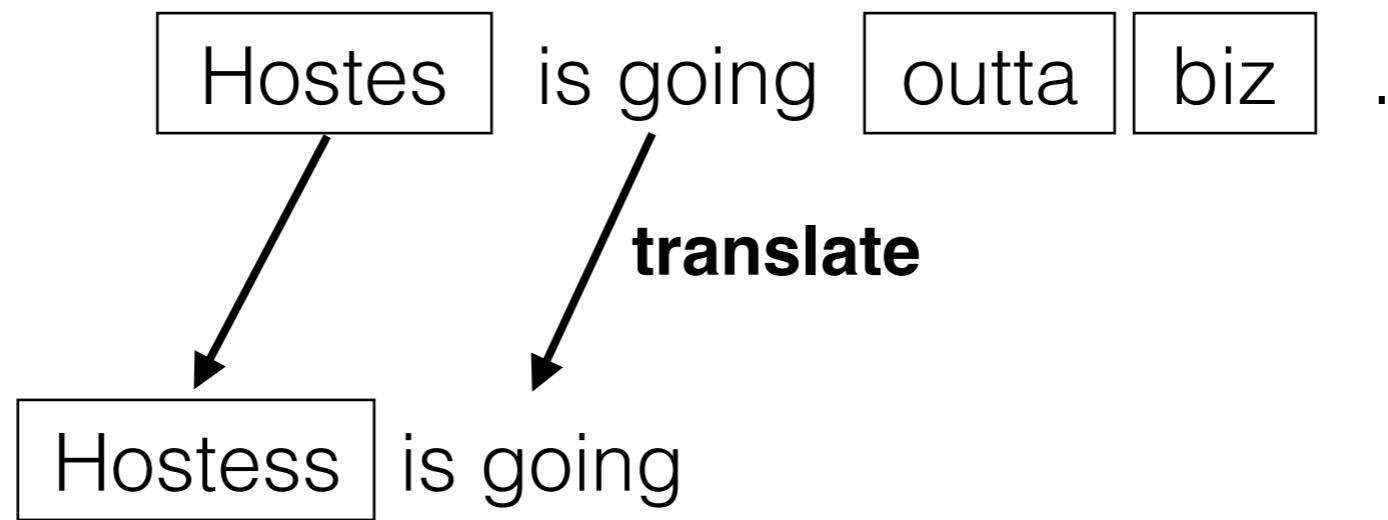
Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao.

“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



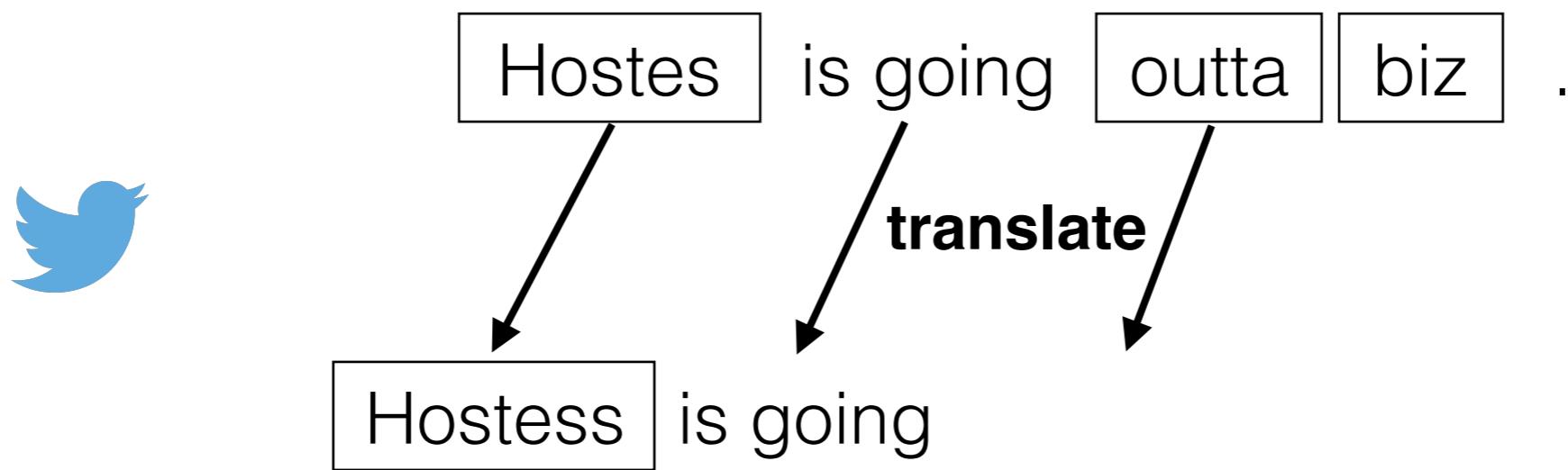
Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao.

“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



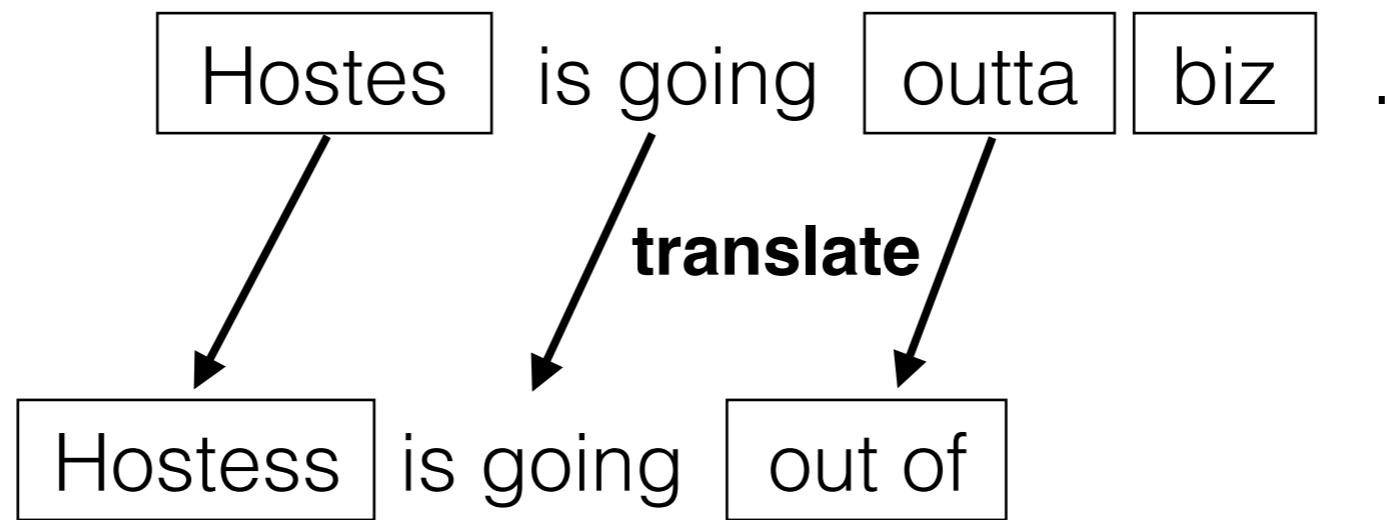
Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao.

“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



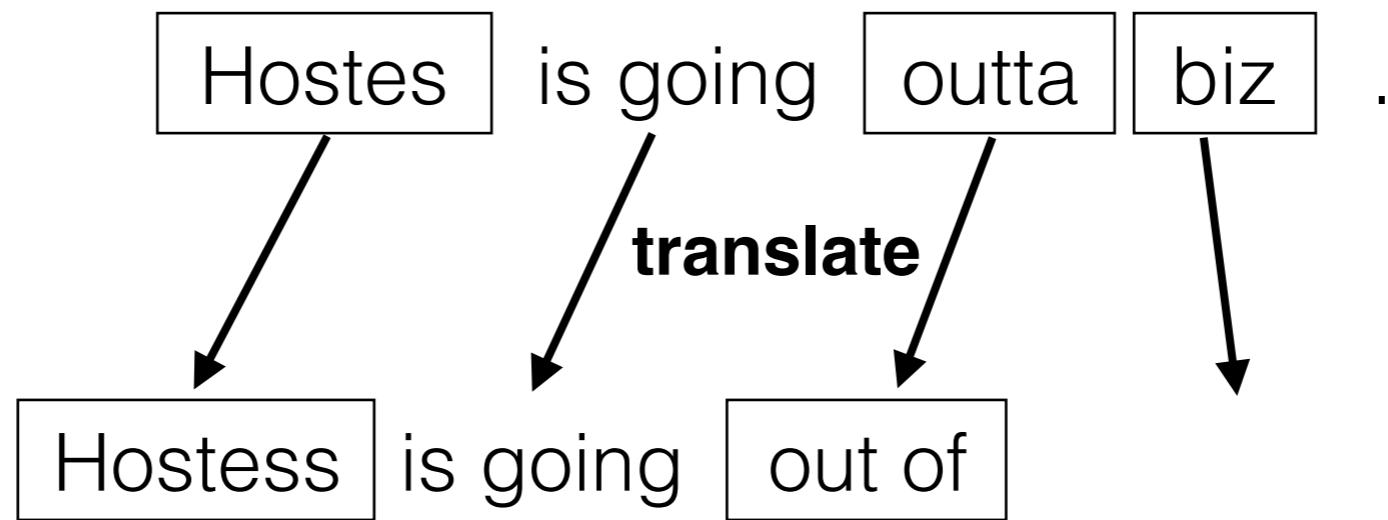
Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao.

“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



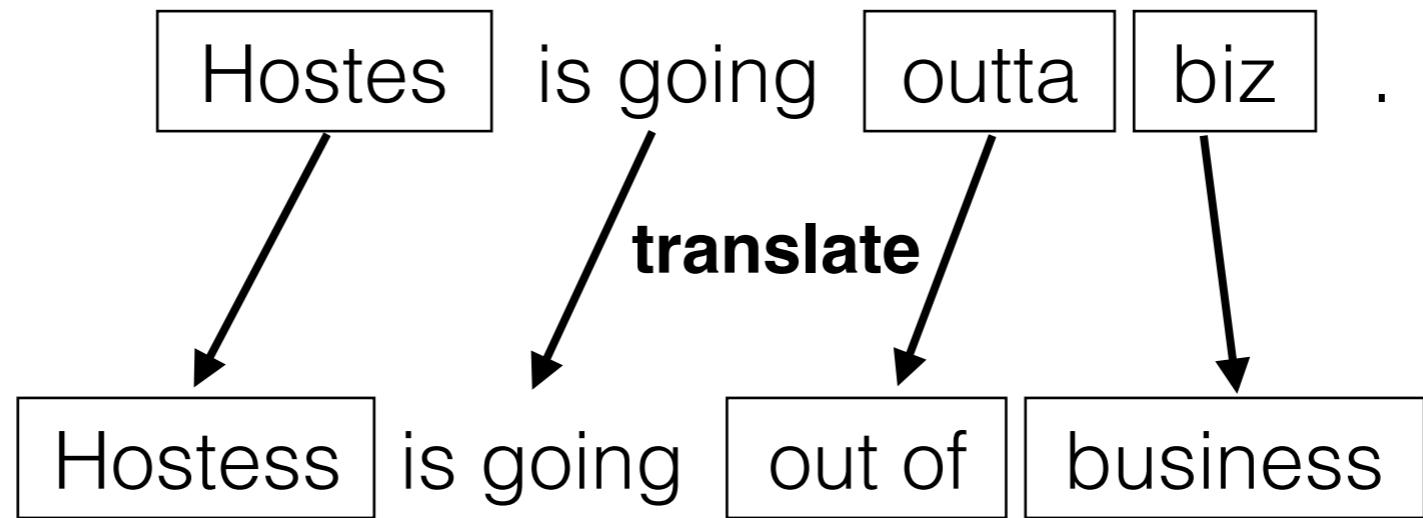
Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao.

“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



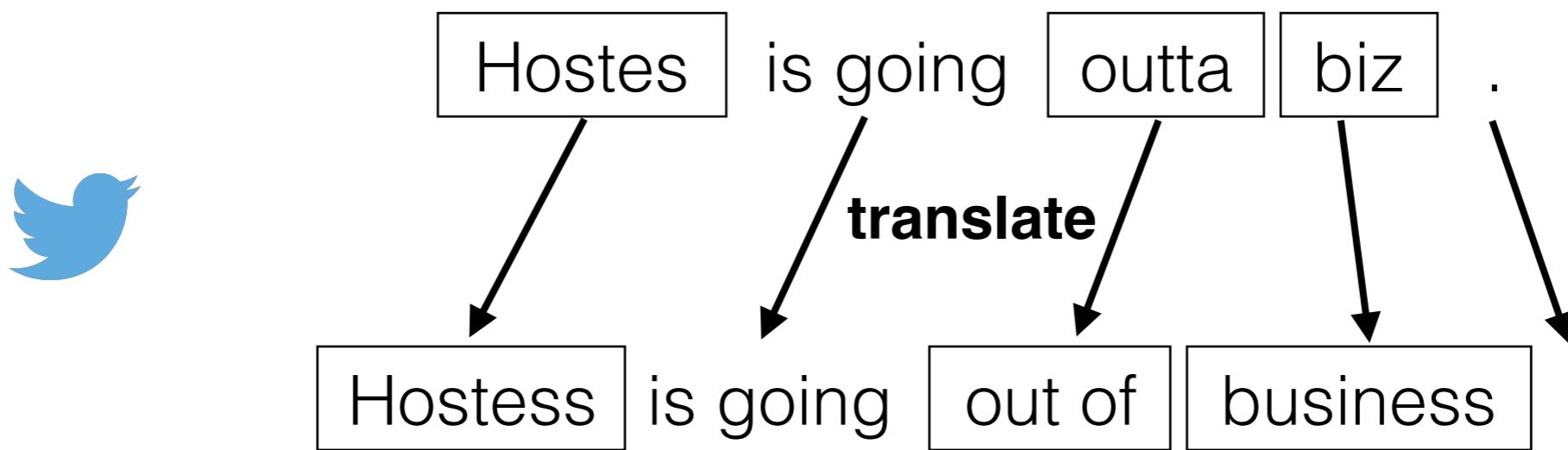
Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao.

“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



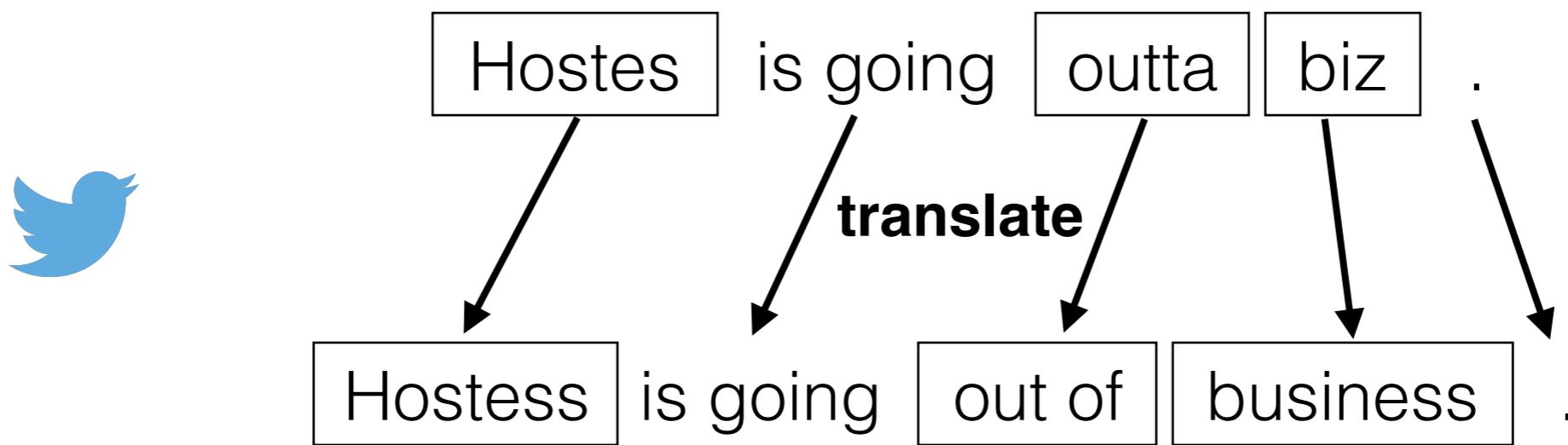
Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao.

“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



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)

# Twitter is a powerful resource

thousands of users  
talk about both big/micro events daily

# Twitter is a powerful resource

thousands of users  
talk about both big/micro events daily



# Twitter is a powerful resource

thousands of users  
talk about both big/micro events daily



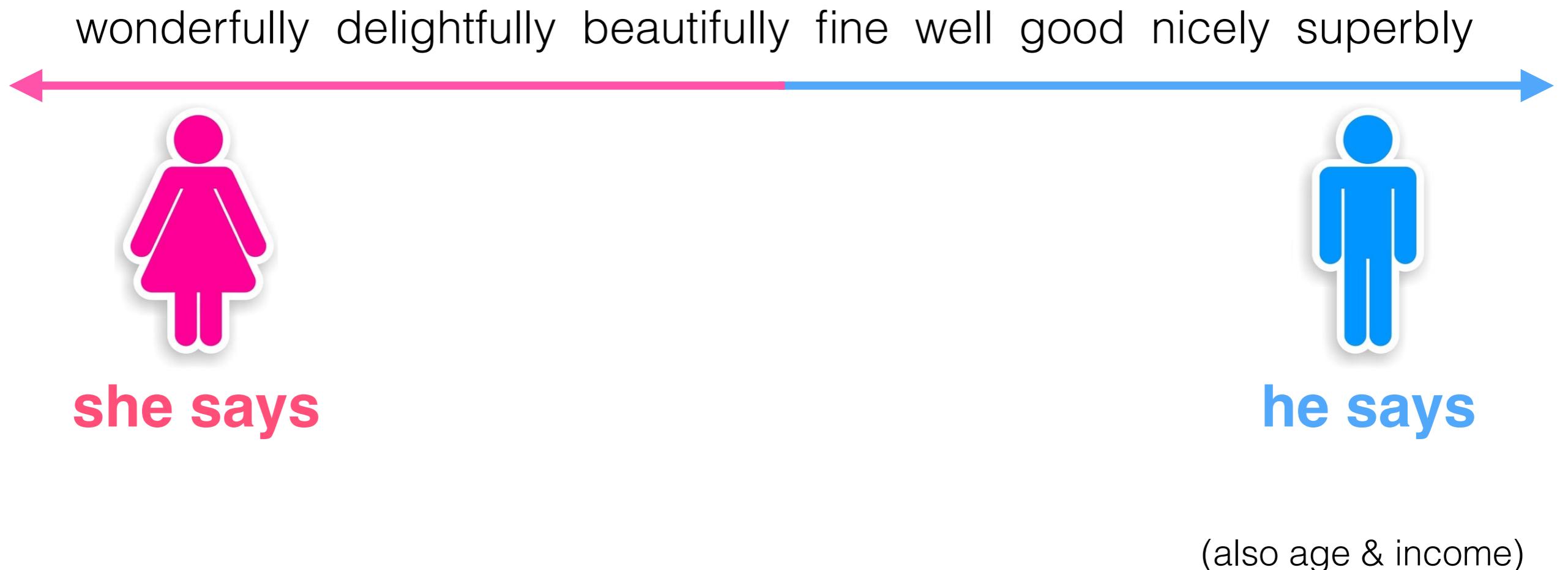
a very broad range of paraphrases:  
synonyms, misspellings, slang, acronyms and colloquialisms

# Diversity in Dialog System



... (21 different paraphrases)

# We speak in style!



# Stylistic NLG via Paraphrasing

	erroneous	→	correct	(Xu et al. EMNLP '11)
	writer style	↔	plain	(Xu et al. COLING '12)
	complex	→	simple	(Xu et al. TACL '15) (Xu et al. TACL '16)
	noisy	→	standard	(Xu et al. BUCC '13) (Xu et al. TACL '14) (Xu et al. SemEval '15)
	feminine	↔	masculine	(Preotiu, Xu, Ungar AAAI '16)

# Stylistic NLG via Paraphrasing



erroneous



correct

(Xu et al. EMNLP '11)



writer style



plain

(Xu et al. COLING '12)



complex



simple

(Xu et al. TACL '15)  
(Xu et al. TACL '16)



noisy



standard

(Xu et al. BUCC '13)  
(Xu et al. TACL '14)  
(Xu et al. SemEval '15)



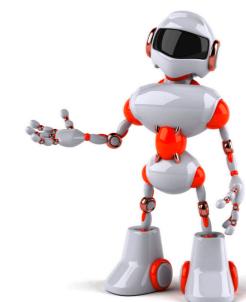
feminine



masculine

(Preotiu, Xu, Ungar AAAI '16)

and more language variations  
(future work) ...





After all, we want a robot that can talk and take selfies ...

# Thank You

thanku

# Thank You

thank u 4 ur time

thankning you

gratitude

appreciate it

thx

tyvm

thanks

3x

say thanks

thank you very much

thnx

thanks a ton

wawwww thankkkkkkkkkkk you alottttttttt!

thanks a lot

I am grateful