Can paraphrase be a ultimate solution for NLU and NLG?

Wei Xu

Department of Computer Science and Engineering

THE OHIO STATE UNIVERSITY

Follow @cocoweixu

Jul-28-2017 @ Google NYC

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

Follow @cocoweixu

Jul-28-2017 @ Google NYC

Images http://www.ufunk.net/en/gadgets/selfie-arm/

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.*

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

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


BOB DYLAN A NOBEL AWARD IN LITERATURE

'CAPITAL OF KITSCH' APPRECIATING ABSURDITY AND ANTIQUITY IN SKOPJE

'IT'S JUST WORDS' THAT'S WHAT TRUMP SAYS. THESE 2 WOMEN DIFFER.

The New York Times

INTERNATIONAL EDITION | FRIDAY, OCTOBER 14, 2016

Trump gets it right about one thing

Nicholas Kristof

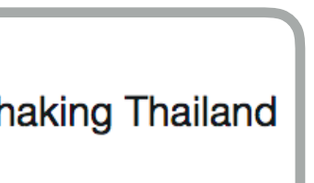
ISIS turning drones into airborne explosives

WASHINGTON

U.S. must now find ways to thwart small aircraft not only spying but killing

Filipinos back tough approach

Despite woes abroad, polls show support for Duterte's harsh measures

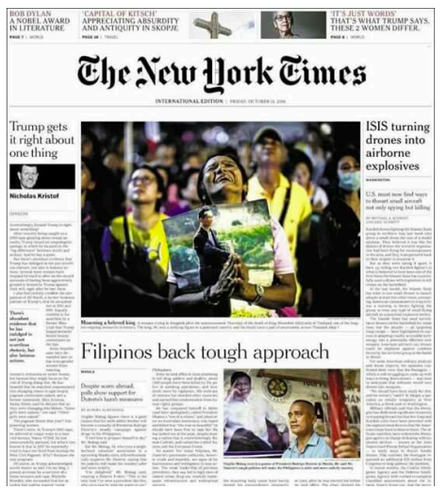

The New York Times @nytimes · 12 Oct 2016

Worries over the health of King Bhumibol Adulyadej are shaking Thailand
nyti.ms/2dRzPcr

5 261 144

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

Web is a Gold Mine



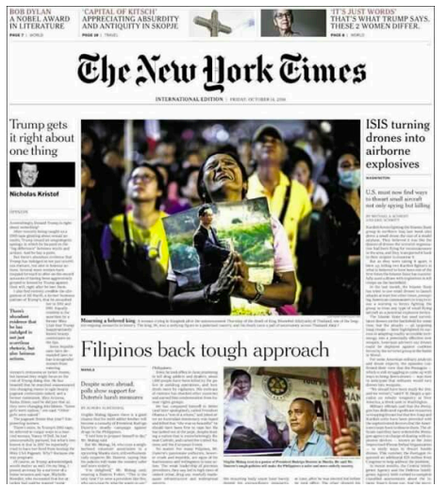
The New York Times @nytimes · 12 Oct 2016
Worries over the health of King Bhumibol Adulyadej are shaking Thailand
nyti.ms/2dRzPcr
5 261 144

Career Synchronicity @careersync_now · 12 Oct 2016
Fears for King's Health Shake Thailand ift.tt/2d7frGd

Paraphrase

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

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
5 261 144

Career Synchronicity @careersync_now · 12 Oct 2016
Fears for King's Health Shake Thailand ift.tt/2d7frGd

Paraphrase

Herbert Buchsbaum @herbertnyt · 12 Oct 2016
New bulletin from Thai palace: King is still on a ventilator and in unstable condition. nyti.ms/2dW1A37

Non-Paraphrase

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

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; and many others]

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)

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

[Twitter Paraphrase Corpus]

51,524 sentence pairs
(manually annotated)

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

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

Timely Paraphrases

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

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

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

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

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 .

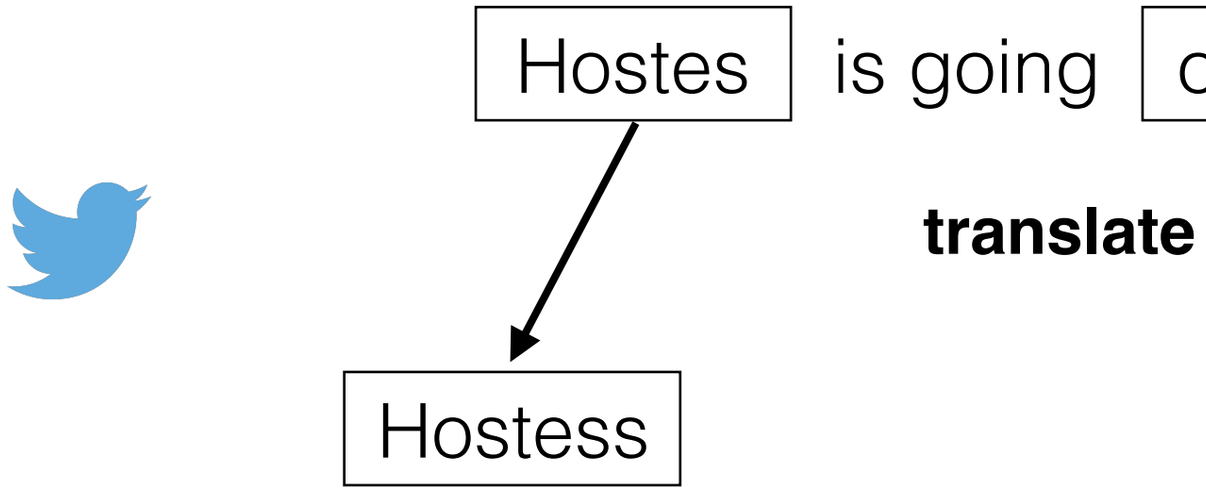
translate

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

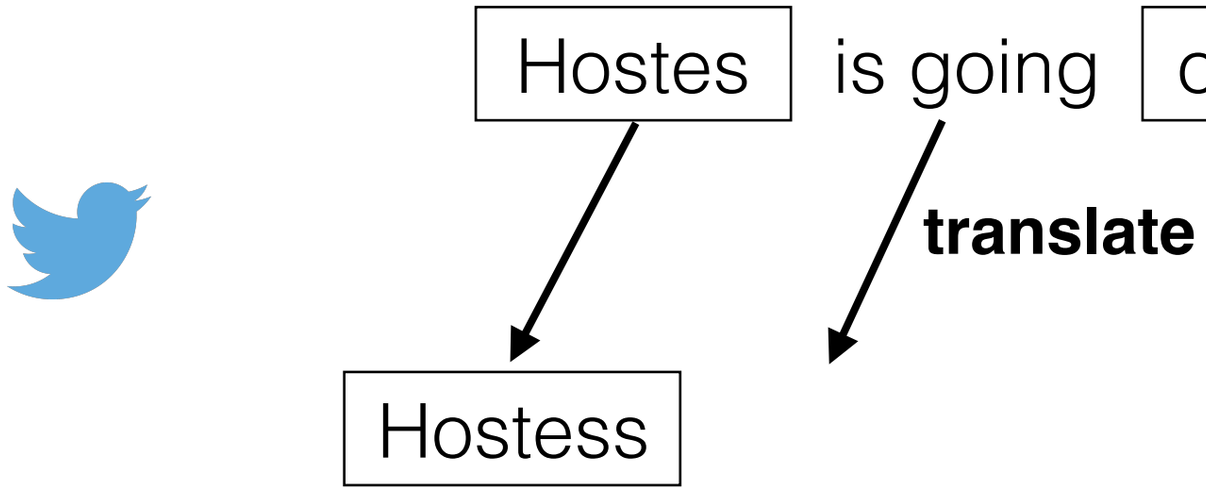


Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao. “Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-gram Models” In EMNLP (2011)

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

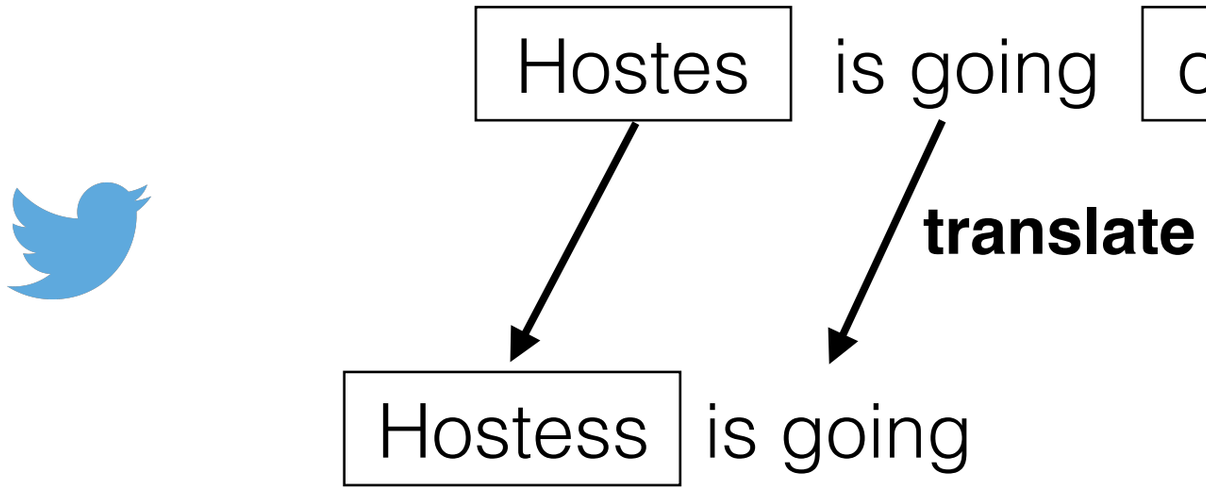


Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao. “Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-gram Models” In EMNLP (2011)

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

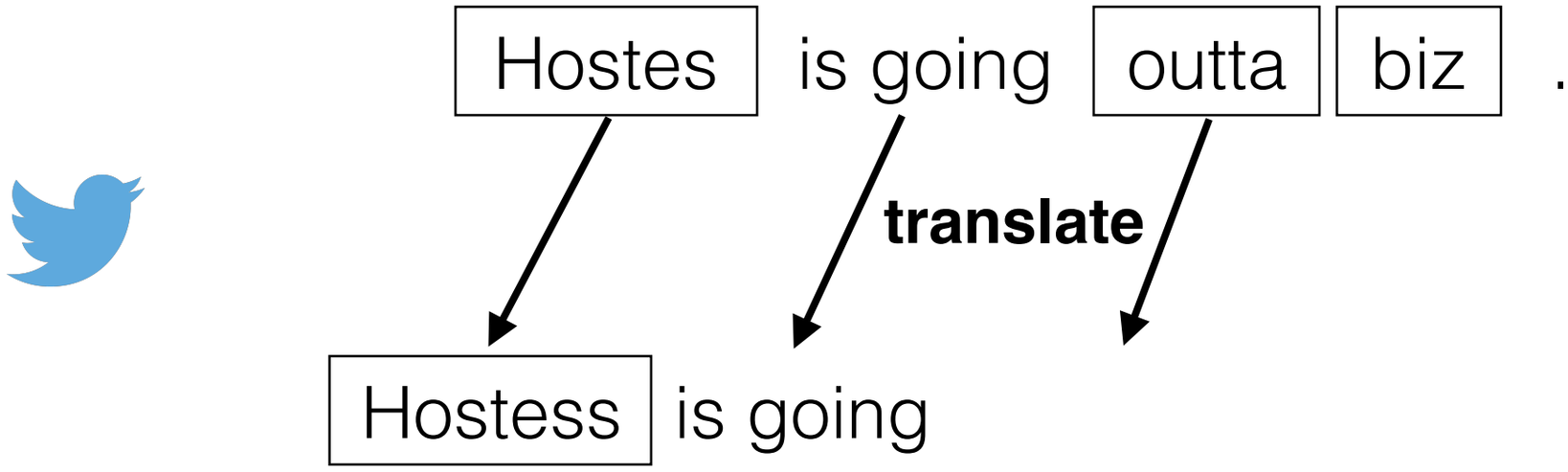


Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao. “Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-gram Models” In EMNLP (2011)

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

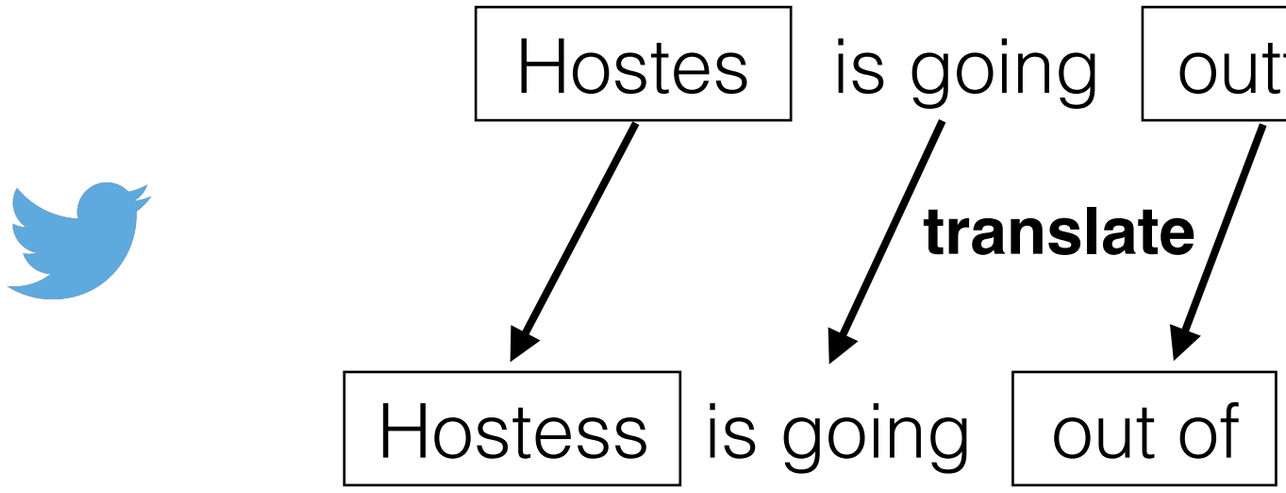


Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao. “Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-gram Models” In EMNLP (2011)

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

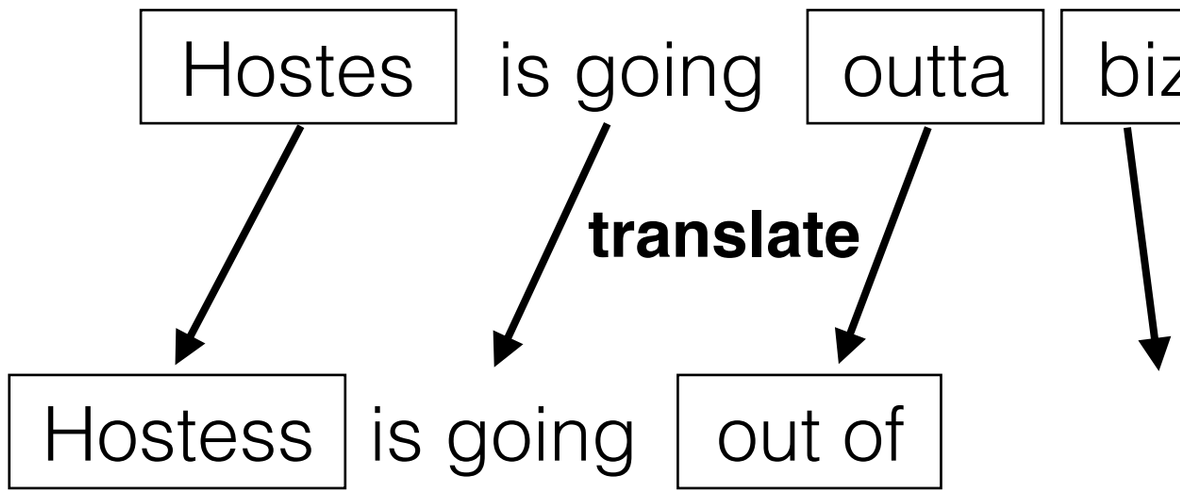


Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao. "Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-gram Models" In EMNLP (2011)

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

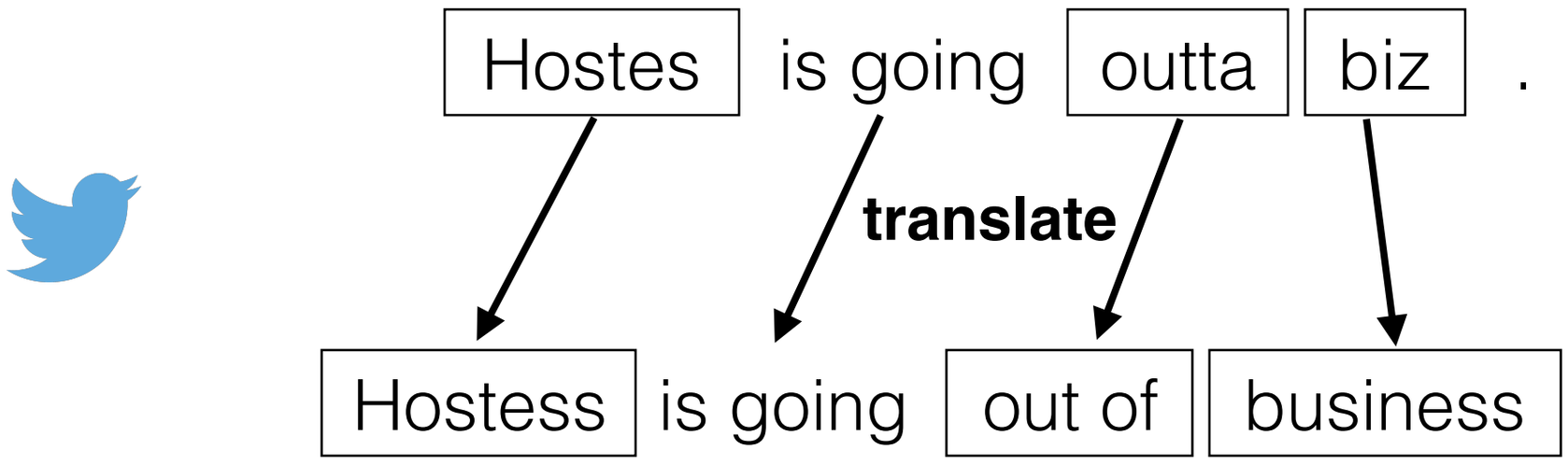
Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, Wei Xu. "Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition" In WNUT (2015)

Noisy Text Normalization



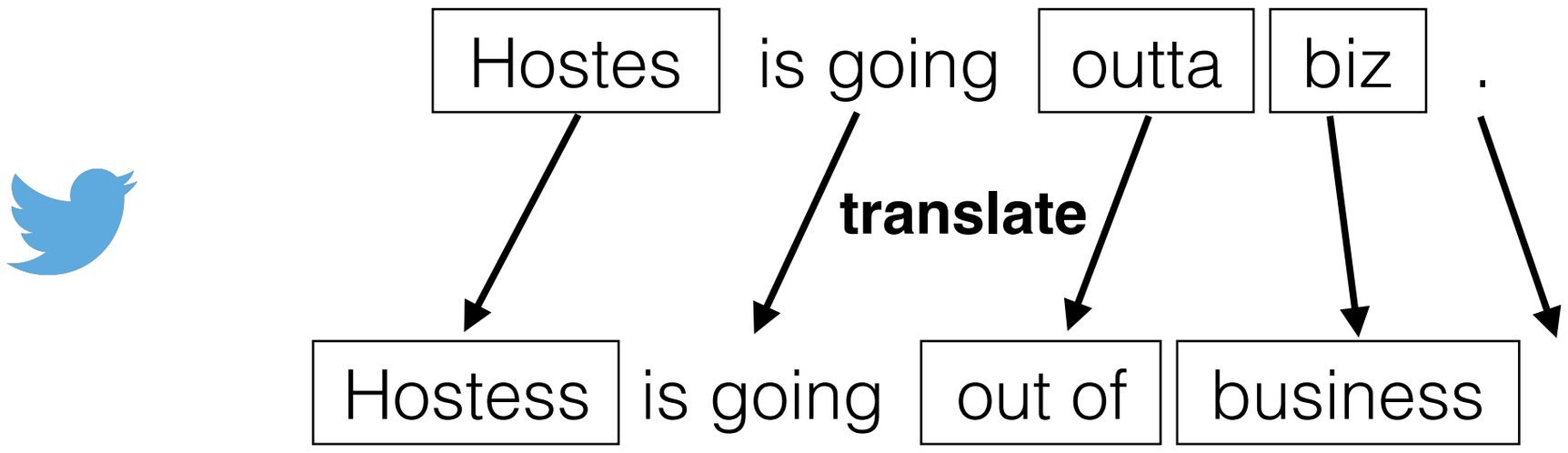
Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao.
“Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-gram Models” In EMNLP (2011)
Wei Xu, Alan Ritter, Ralph Grishman. “Gathering and Generating Paraphrases from Twitter with Application to Normalization” 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



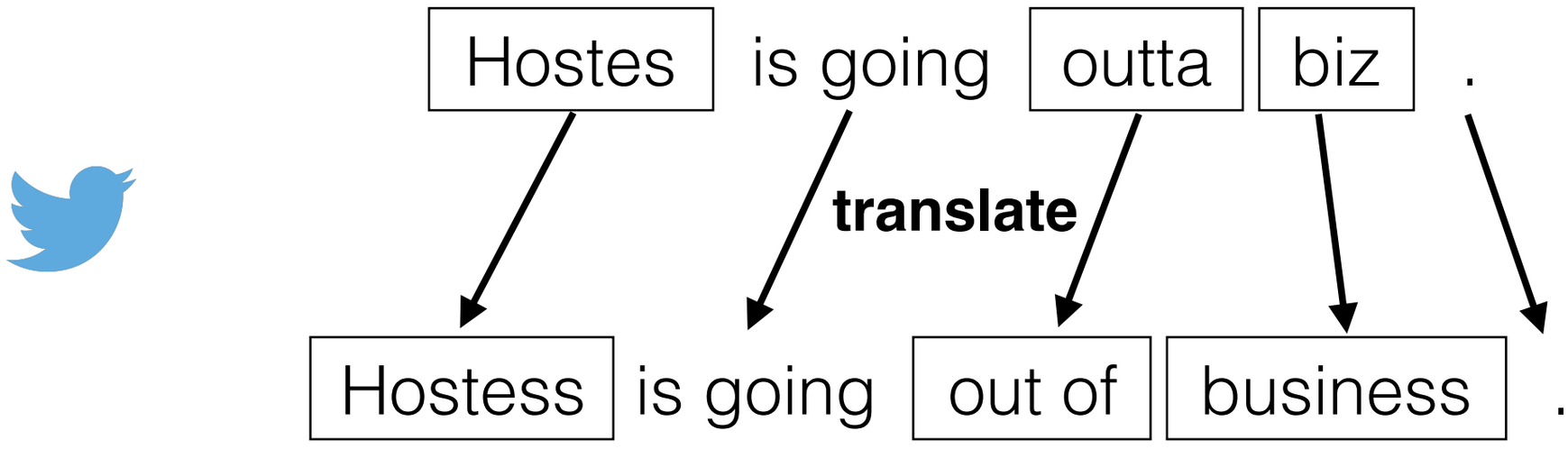
Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, Le Zhao.
“Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-gram Models” In EMNLP (2011)
Wei Xu, Alan Ritter, Ralph Grishman. “Gathering and Generating Paraphrases from Twitter with Application to Normalization” 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

Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. “Extracting Lexically Divergent Paraphrases from Twitter” In TACL (2014)
Wei Xu, Chris Callison-Burch, Bill Dolan. “SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter” In SemEval (2015)

Twitter is a powerful resource

thousands of users
talk about both big/micro events daily

Very diverse!

Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. “Extracting Lexically Divergent Paraphrases from Twitter” In TACL (2014)
Wei Xu, Chris Callison-Burch, Bill Dolan. “SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter” In SemEval (2015)

Twitter is a powerful resource

thousands of users
talk about both big/micro events daily

Very diverse!

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

Diversity in Dialog System

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?

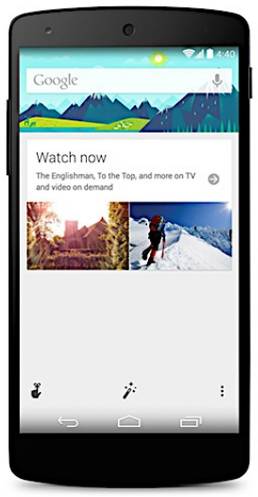
who else wants to get a beer?

trying to get a beer?

… (21 different paraphrases)

Apple Siri

Google Now



Windows Cortana

We speak in style!

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)

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 **(Preotiuc, Xu, Ungar AAAI ’16)**

Wei Xu. “From Shakespeare to Twitter: What are Language Styles all about? ” in StyVa@EMNLP (2017)

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 **(Preotiuc, Xu, Ungar AAAI ’16)**

and more language variations (future work) …

Wei Xu. “From Shakespeare to Twitter: What are Language Styles all about? ” in StyVa@EMNLP (2017)

Image http://www.taipeitimes.com/News/biz/photo/2015/02/23/2008113120

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

Thank You

Follow @cocoweixu

http://web.cse.ohio-state.edu/~weixu/

thank u 4 ur time

thanku

Thank You

thanking you

appreciate it

gratitude

thx

3x

tyvm

thanks

say thanks

thank you very much

thnx

thanks a ton

wawwww thankkkkkkkkkkk you alottttttttttt!

thanks a lot

I am grateful

Follow @cocoweixu

http://web.cse.ohio-state.edu/~weixu/