From Shakespeare to Twitter:

What are Language Styles all about?

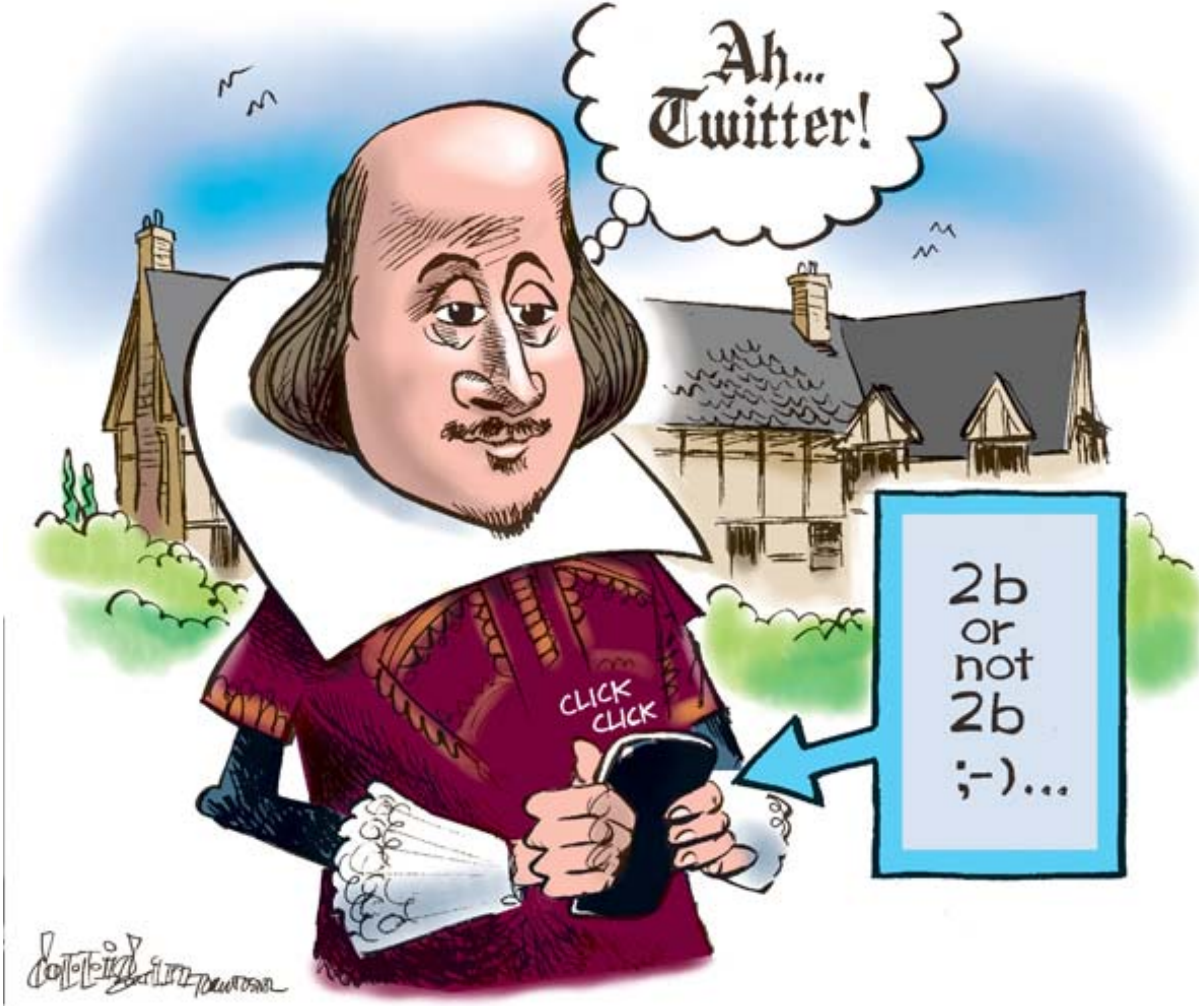
Wei Xu

Department of Computer Science and Engineering

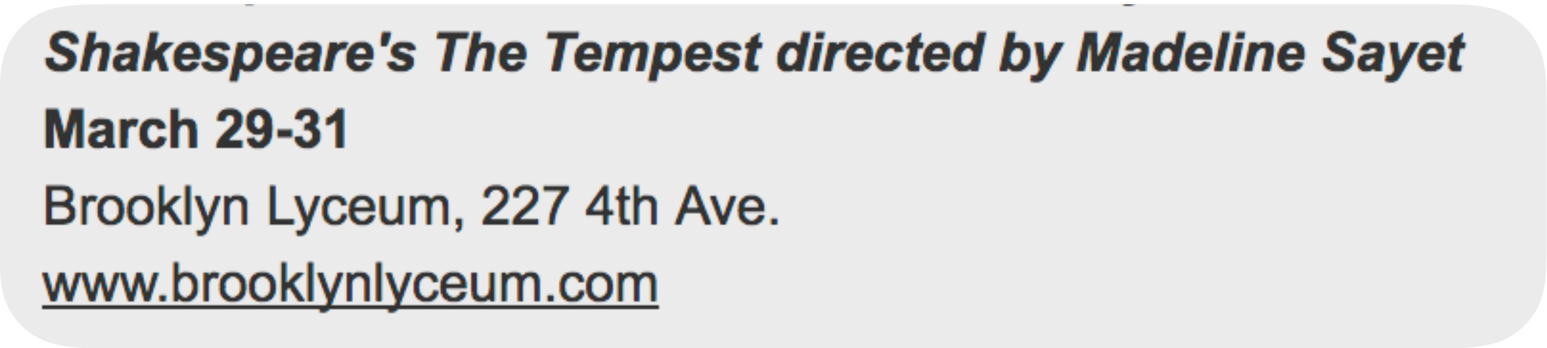
The Ohio State University

Follow @cocoweixu

Sep-08-2017 @ EMNLP 2017


Ah... Twitter!
CLICK CLICK
2b or not 2b ;-)...

My very first Shakespeare play



Shakespeare's The Tempest directed by Madeline Sayet

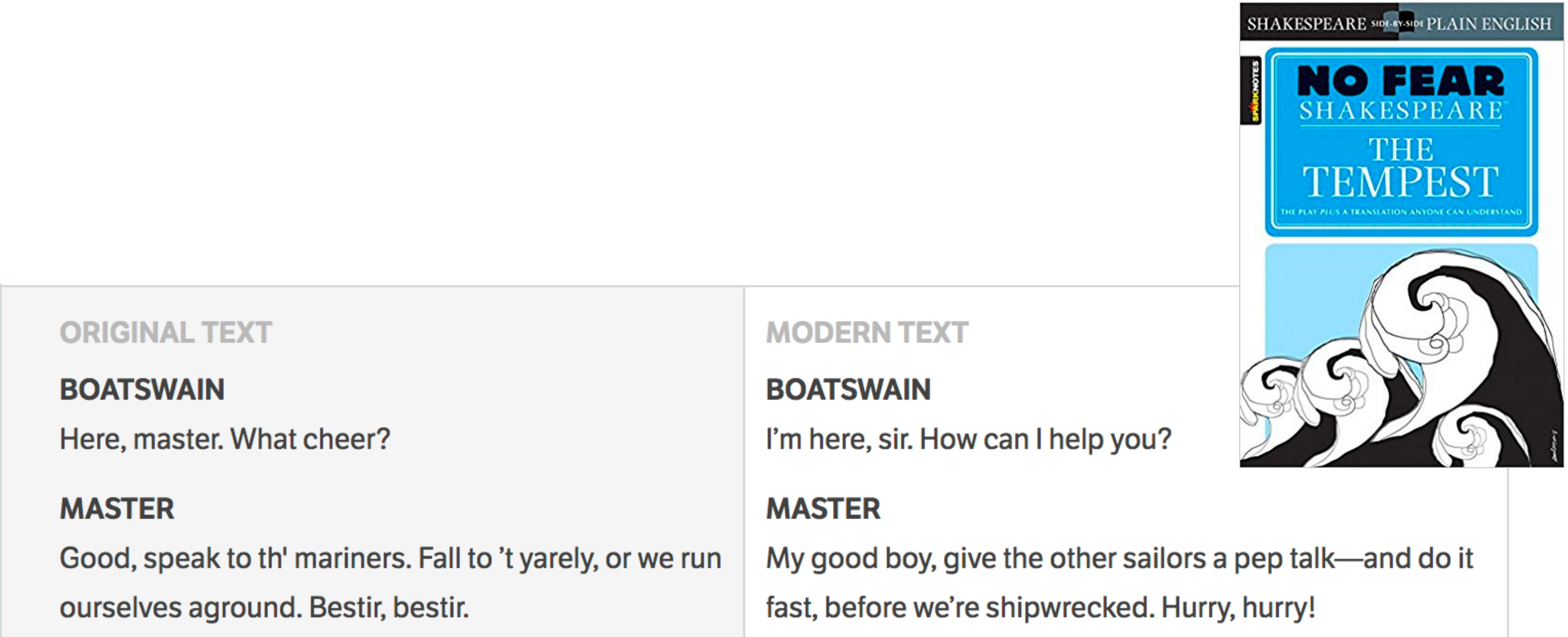
March 29-31

Brooklyn Lyceum, 227 4th Ave.

www.brooklynlyceum.com

(Brooklyn - March 31, 2012)

How can I learn Early Modern English in no time?


SHAKESPEARE SIDE-BY-SIDE PLAIN ENGLISH
SPARKNOTES
NO FEAR
SHAKESPEARE
THE TEMPEST
THE PLAY PLUS A TRANSLATION ANYONE CAN UNDERSTAND

ORIGINAL TEXT	MODERN TEXT
BOATSWAIN	**BOATSWAIN**
Here, master. What cheer?	I'm here, sir. How can I help you?
MASTER	**MASTER**
Good, speak to th' mariners. Fall to 't yarely, or we run ourselves aground. Bestir, bestir.	My good boy, give the other sailors a pep talk—and do it fast, before we're shipwrecked. Hurry, hurry!

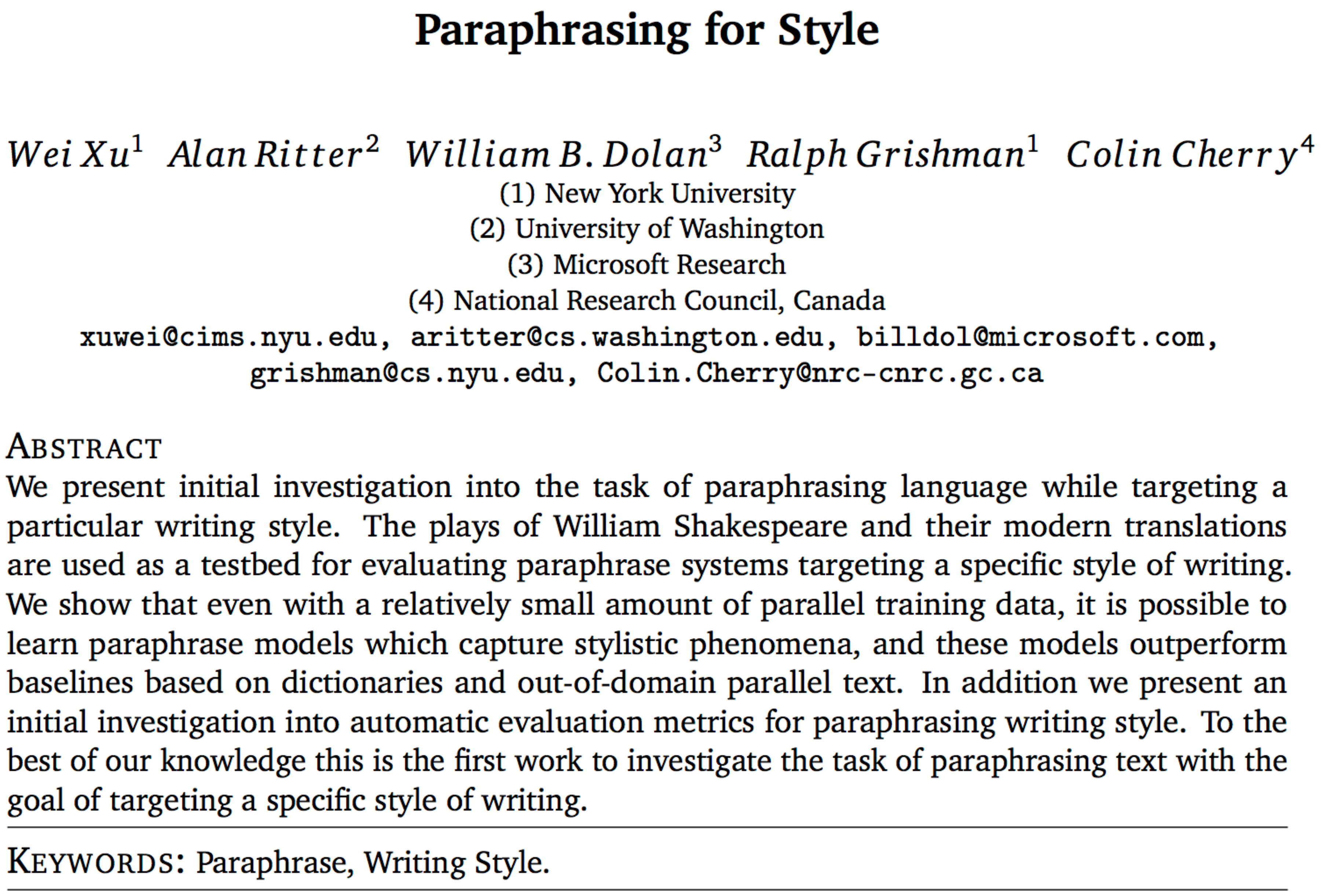
I trained machine translation models …

Shakespearized (real system output):

If you will not be turn’d, you will be undone!

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry. “Paraphrasing for Style” (COLING 2012)

My very first paper on Stylistics



Paraphrasing for Style

Wei Xu[1] *Alan Ritter*[2] *William B. Dolan*[3] *Ralph Grishman*[1] *Colin Cherry*[4]
(1) New York University
(2) University of Washington
(3) Microsoft Research
(4) National Research Council, Canada
xuwei@cims.nyu.edu, aritter@cs.washington.edu, billdol@microsoft.com,
grishman@cs.nyu.edu, Colin.Cherry@nrc-cnrc.gc.ca

ABSTRACT

We present initial investigation into the task of paraphrasing language while targeting a particular writing style. The plays of William Shakespeare and their modern translations are used as a testbed for evaluating paraphrase systems targeting a specific style of writing. We show that even with a relatively small amount of parallel training data, it is possible to learn paraphrase models which capture stylistic phenomena, and these models outperform baselines based on dictionaries and out-of-domain parallel text. In addition we present an initial investigation into automatic evaluation metrics for paraphrasing writing style. To the best of our knowledge this is the first work to investigate the task of paraphrasing text with the goal of targeting a specific style of writing.

KEYWORDS: Paraphrase, Writing Style.

(COLING - June 15, 2012)

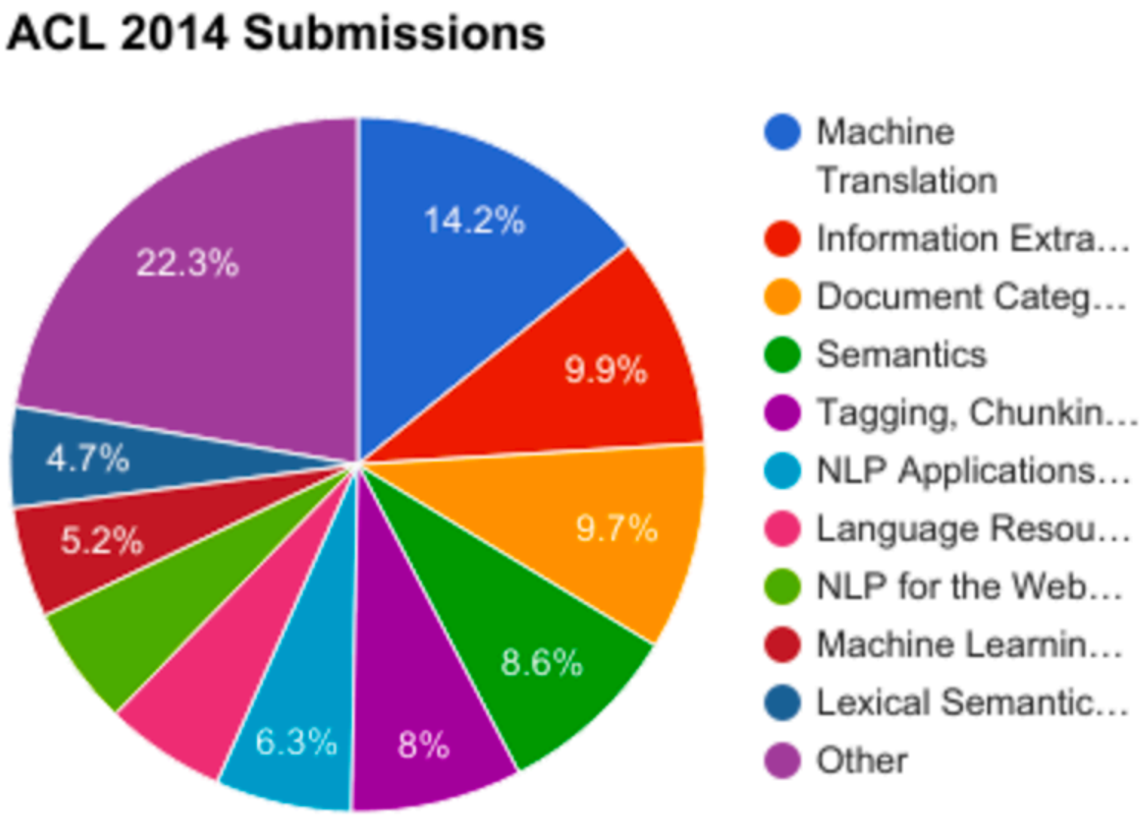
Stylistic Paraphrase becomes a thing!

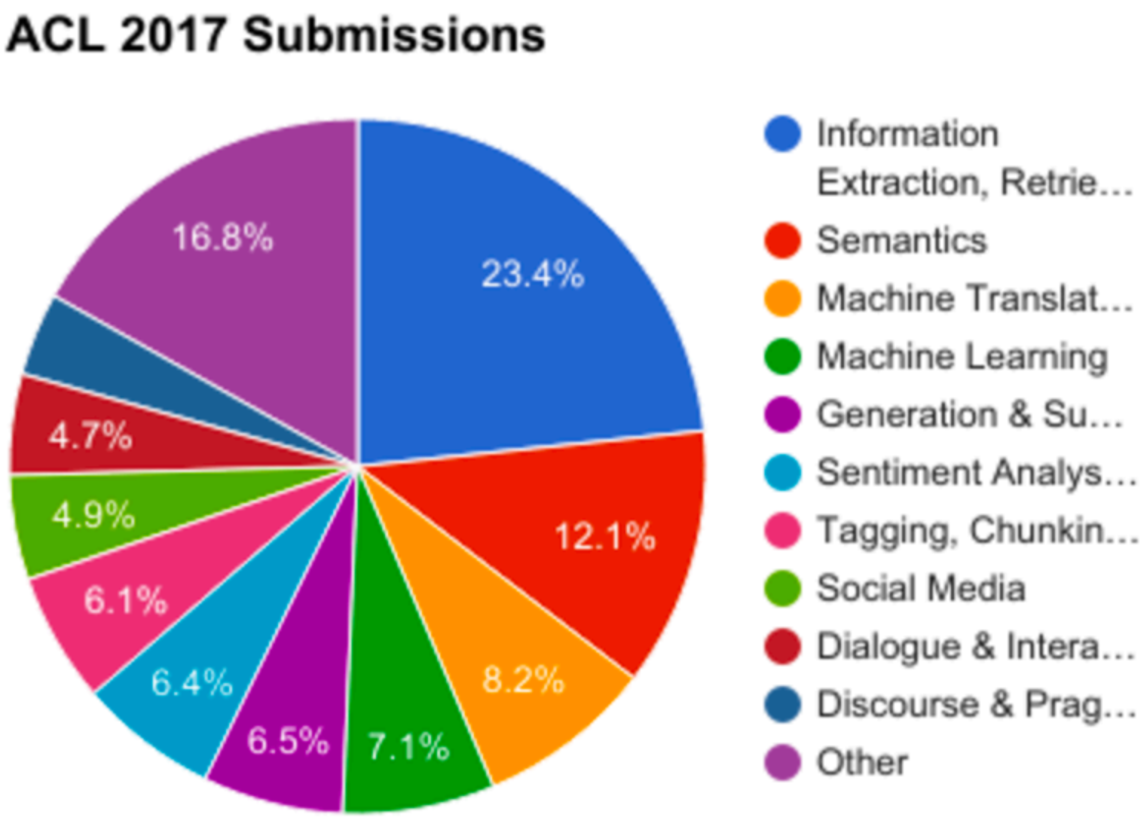
- **Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, Colin Cherry. “Paraphrasing for Style” (COLING 2012)**
- Had Kabbara, Jackie Cheung. “Stylistic Transfer in Natural Language Generation Systems Using Recurrent Neural Networks” (EMNLP Uphill Battles 2016)
- Harsh Jhamtani, Varun Gangal, Eduard Hovy, Eric Nyberg “Shakespearizing Modern Language Using Copy-Enriched Sequence to Sequence Models” (**EMNLP StyVa 2017**)
- Se Won Jang, Jesik Min, Mark Kwon. “Writing Style Conversion using Neural Machine Translation” (Stanford CS224n 2017)

Paraphrasing ≈ Monolingual Translation ≈ T2T Generation

text-to-text

Stylistic Generation becomes a thing!





Generation & Summarization is now in top 5 areas,
while in 2014 it didn't even make top 10

source: https://chairs-blog.acl2017.org/

Can help User-generated Text

Hostes is going outta biz.

Normalized:

Hostess is going out of business.

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

Wei Xu, Alan Ritter, Ralph Grishman. "Gathering and Generating Paraphrases from Twitter with Application to Normalization" (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" (WNUT 2015)

Can help children read

News article:

Slightly more fourth-graders nationwide are reading proficiently compared with a decade ago.

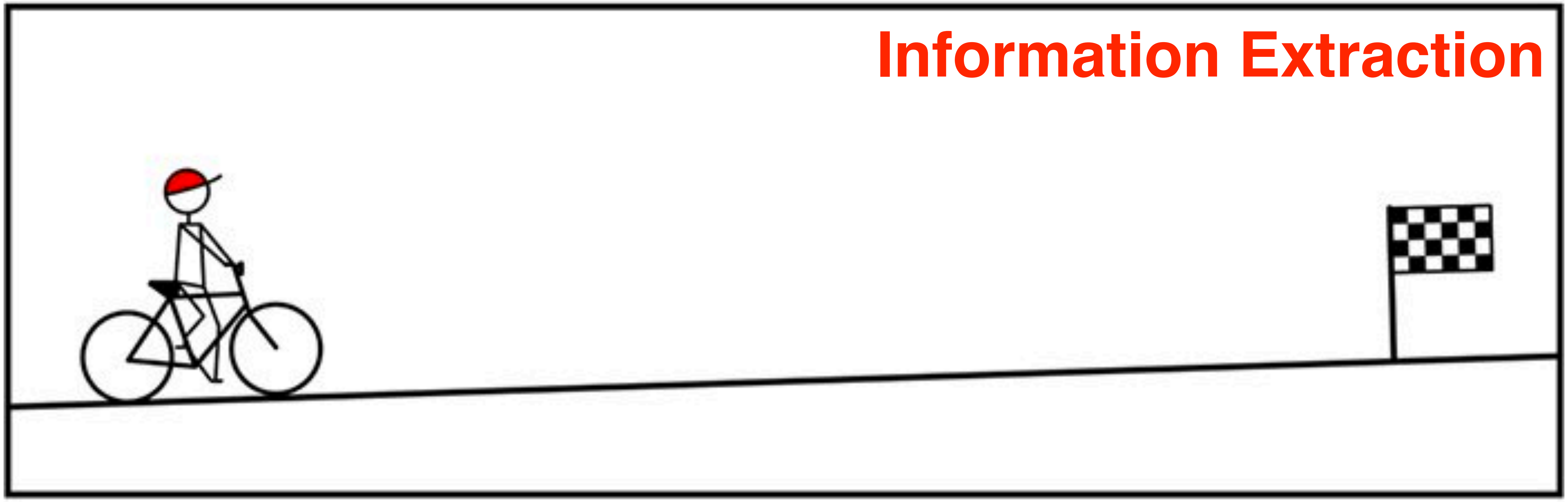
Simplified:

Most fourth-graders are better readers than they were 10 years ago.

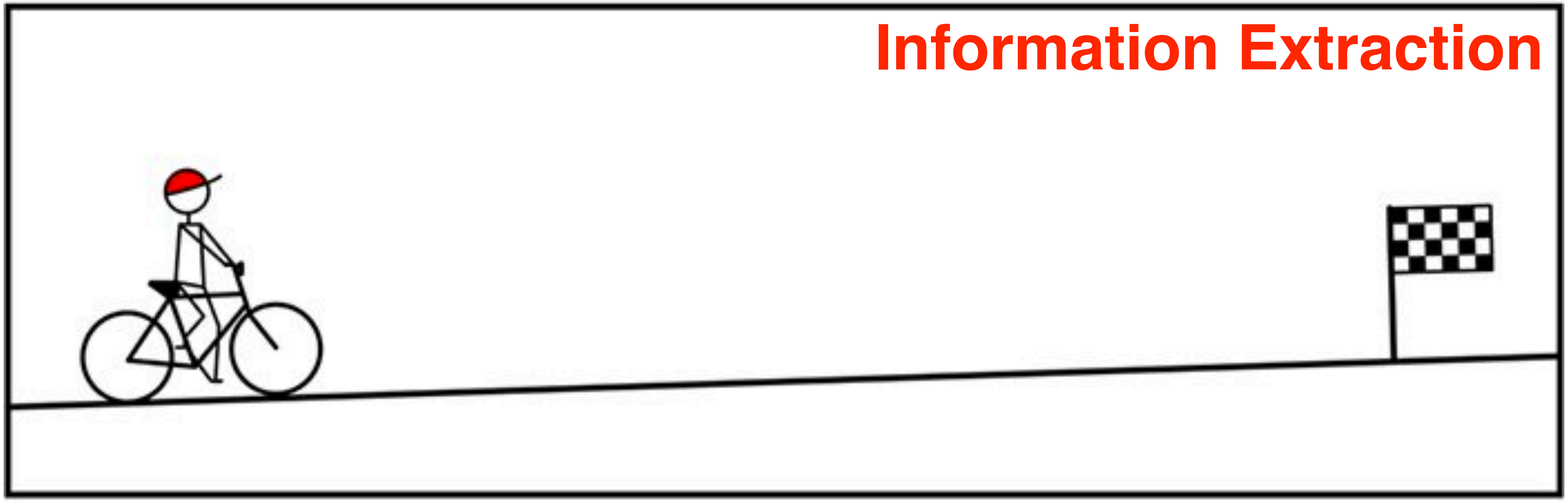
Wei Xu, Chris Callison-Burch, Courtney Napoles “Problems in Current Text Simplification Research: New Data Can Help” (TACL 2015)
Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, Chris Callison-Burch. "Optimizing Statistical Machine Translation for Text Simplification" (TACL 2016)

My Research

My plan



My plan



Information Extraction

Reality

Paraphrase/Stylistics is exciting!
(with many pitfalls)

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

My Research on Stylistics/Paraphrase

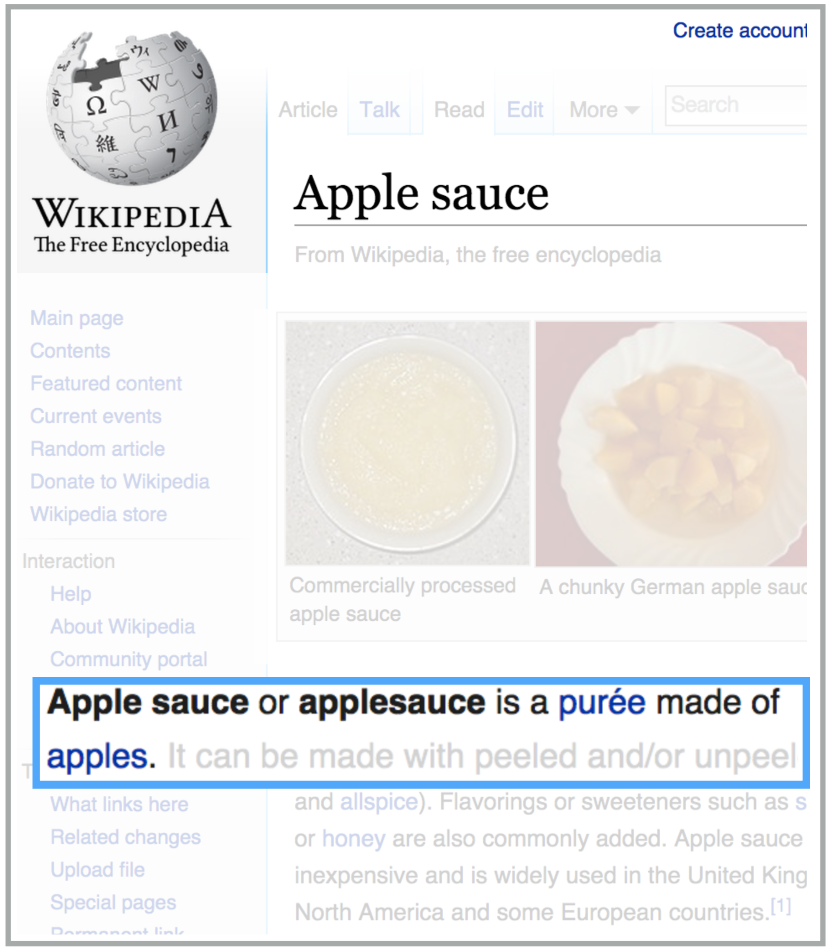
	erroneous	→	correct	(Xu et al. EMNLP ’11)
	writer style	↔	plain	(Xu et al. COLING ’12)
	complex	→	simple	(Xu et al. TACL ’15) (Xu et al. TACL ’16)
	noisy	↔	standard	(Xu et al. BUCC ’13) (Xu et al. TACL ’14) (Xu et al. SemEval ’15) (Lan, Qiu, He, Xu EMNLP ‘ 17)
	feminine	↔	masculine	(Preotiuc, Xu, Ungar AAAI ’16)

Common Pitfalls and Lessons I learned

Lesson #1: Don’t take data quality for granted.

Text Simplification Dataset

Parallel Simple-Normal Wikipedia Corpus was Benchmark since 2010.



Apple sauce

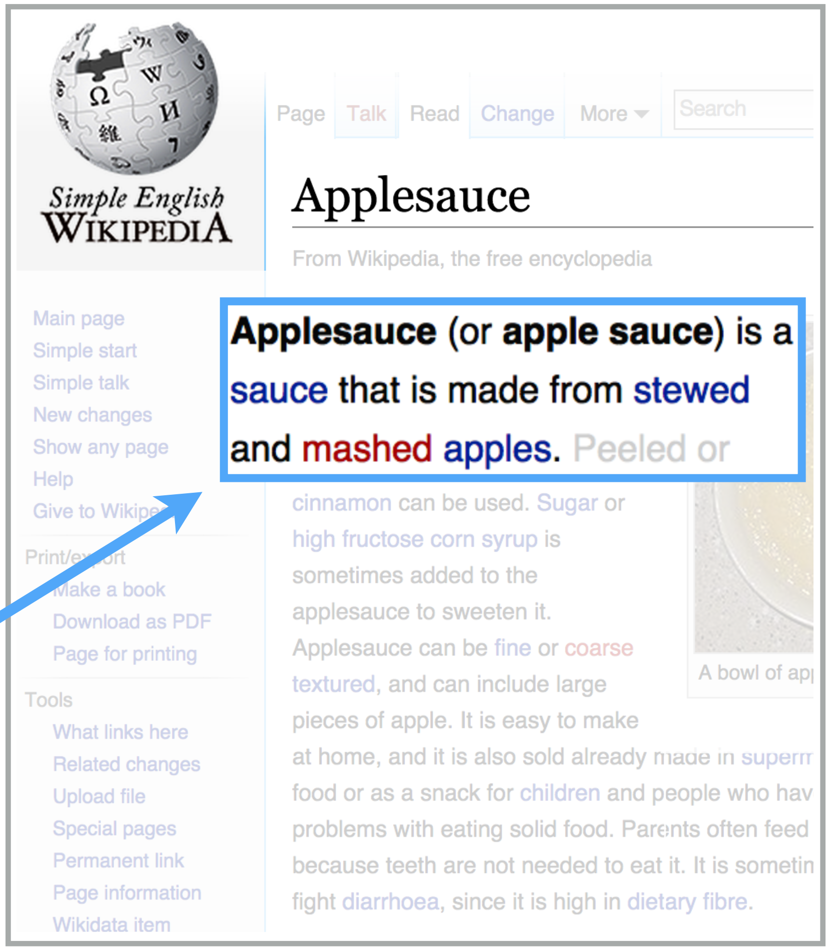
From Wikipedia, the free encyclopedia

Commercially processed apple sauce

A chunky German apple sauc

Apple sauce or **applesauce** is a purée made of apples. It can be made with peeled and/or unpeel

and allspice). Flavorings or sweeteners such as s or honey are also commonly added. Apple sauce inexpensive and is widely used in the United King North America and some European countries.[1]



Applesauce

From Wikipedia, the free encyclopedia

Applesauce (or **apple sauce**) is a sauce that is made from stewed and mashed apples. Peeled or

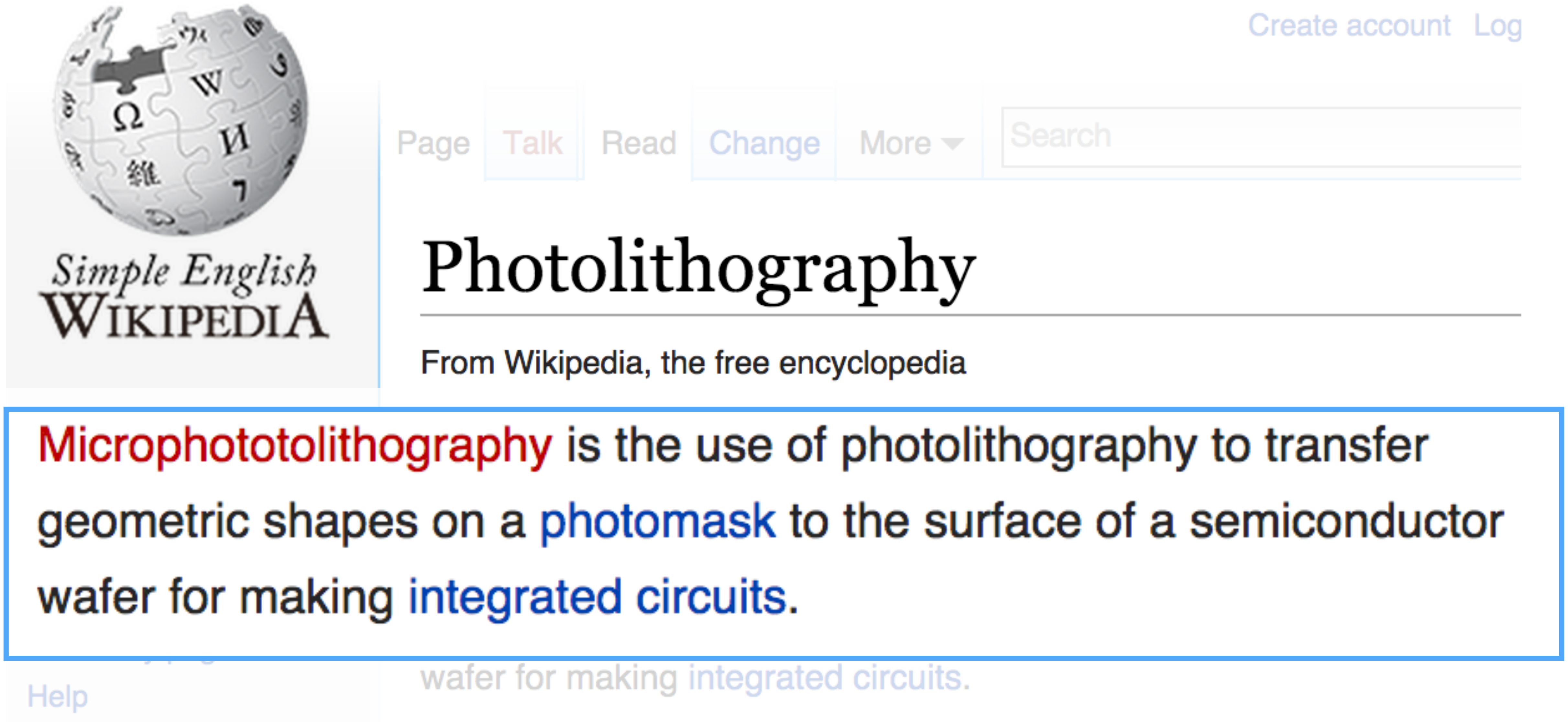
cinnamon can be used. Sugar or high fructose corn syrup is sometimes added to the applesauce to sweeten it. Applesauce can be fine or coarse textured, and can include large pieces of apple. It is easy to make at home, and it is also sold already made in superm food or as a snack for children and people who hav problems with eating solid food. Parents often feed because teeth are not needed to eat it. It is sometin fight diarrhoea, since it is high in dietary fibre.

A bowl of ap

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” (TACL 2015)

Text Simplification Dataset

Pitfall: Simple Wikipedia is not simple.



Photolithography

From Wikipedia, the free encyclopedia

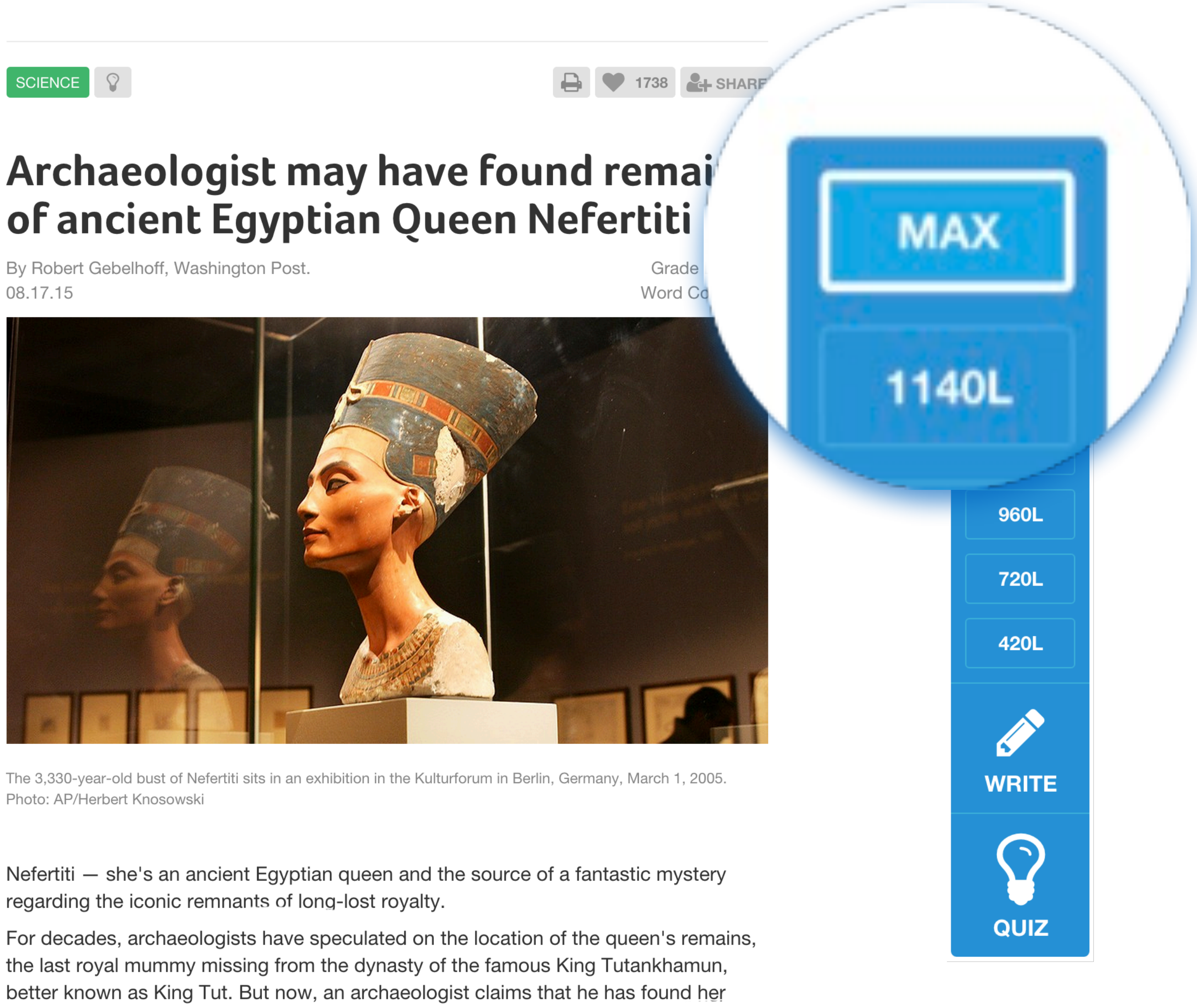
Microphototolithography is the use of photolithography to transfer geometric shapes on a photomask to the surface of a semiconductor wafer for making integrated circuits.

wafer for making integrated circuits.

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” (TACL 2015)

Newsela Corpus

We proposed a new high-quality dataset of news articles simplified by professional editors.



Archaeologist may have found remai
of ancient Egyptian Queen Nefertiti

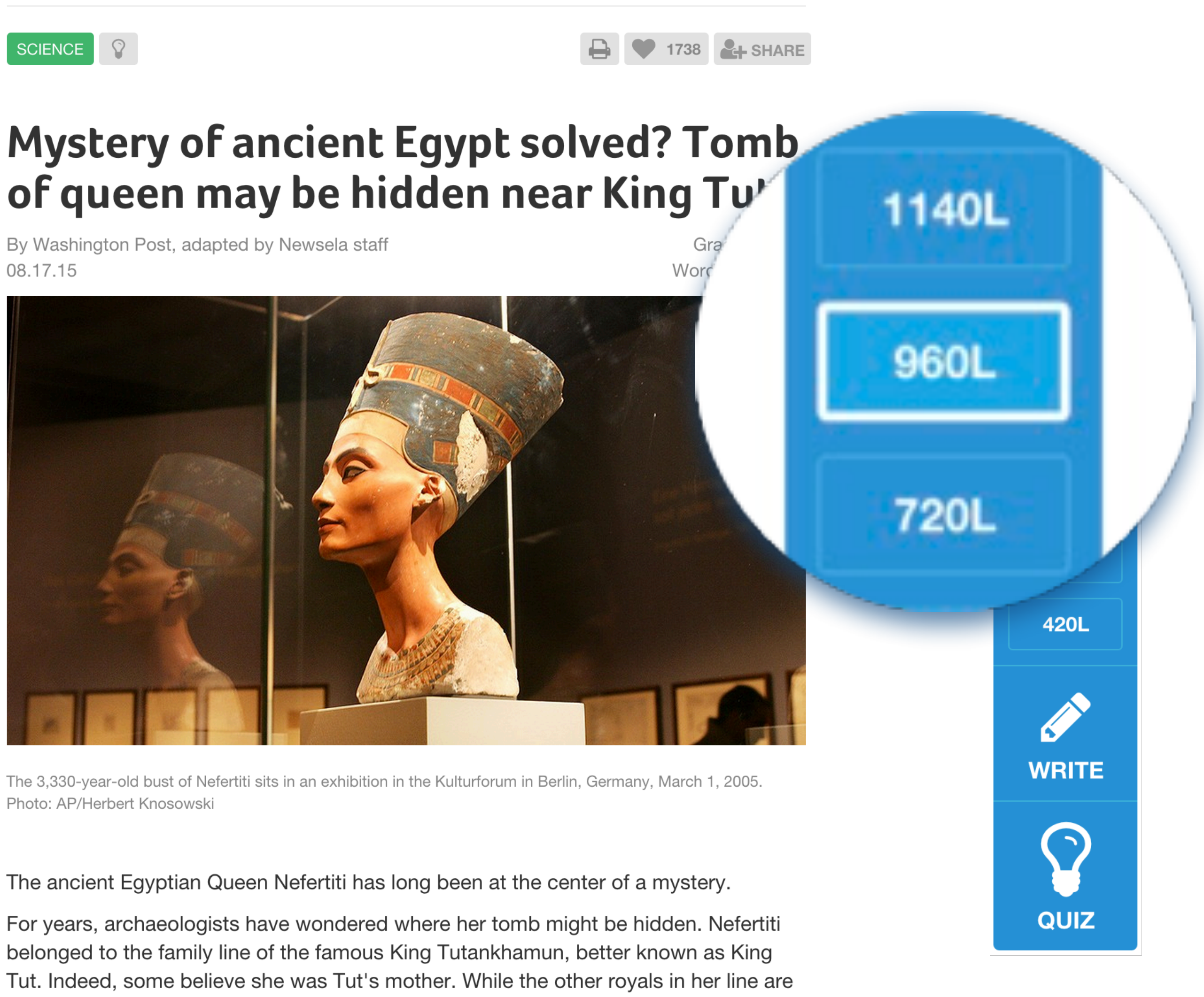
By Robert Gebelhoff, Washington Post.
08.17.15

Grade
Word Co

The 3,330-year-old bust of Nefertiti sits in an exhibition in the Kulturforum in Berlin, Germany, March 1, 2005.
Photo: AP/Herbert Knosowski

Nefertiti — she's an ancient Egyptian queen and the source of a fantastic mystery regarding the iconic remnants of long-lost royalty.

For decades, archaeologists have speculated on the location of the queen's remains, the last royal mummy missing from the dynasty of the famous King Tutankhamun, better known as King Tut. But now, an archaeologist claims that he has found her



Mystery of ancient Egypt solved? Tomb
of queen may be hidden near King Tu

By Washington Post, adapted by Newsela staff
08.17.15

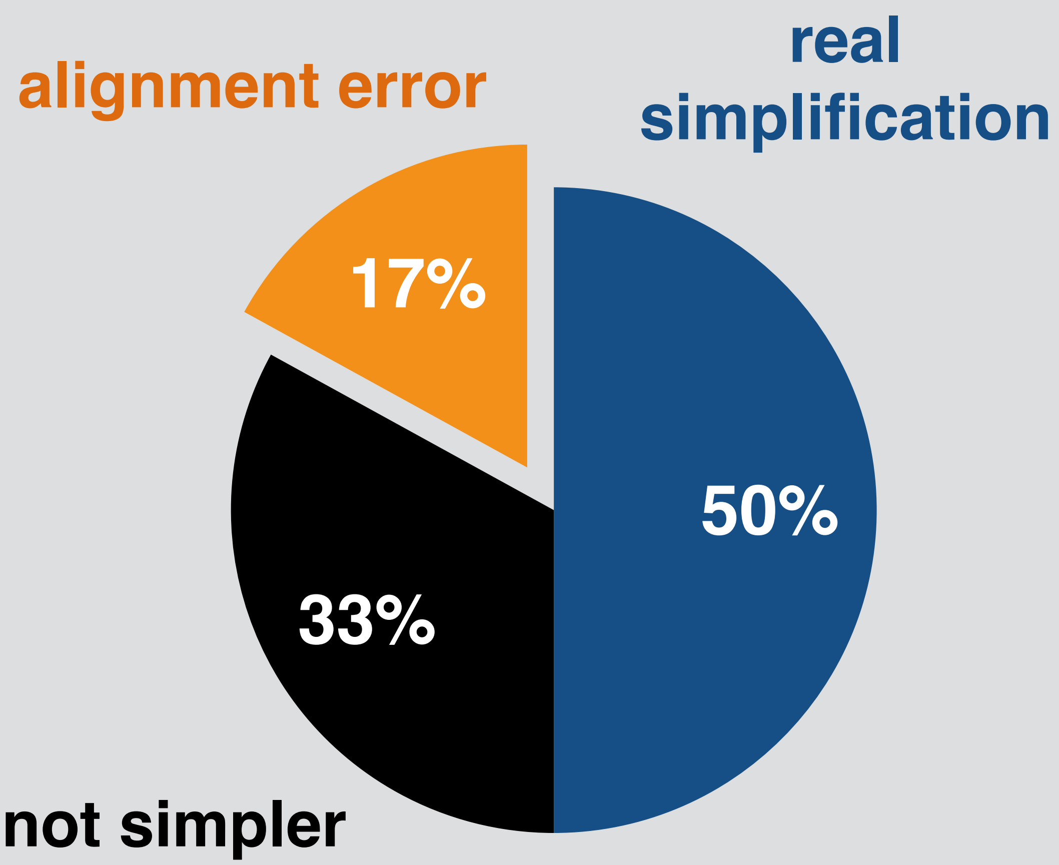
Gra
Word

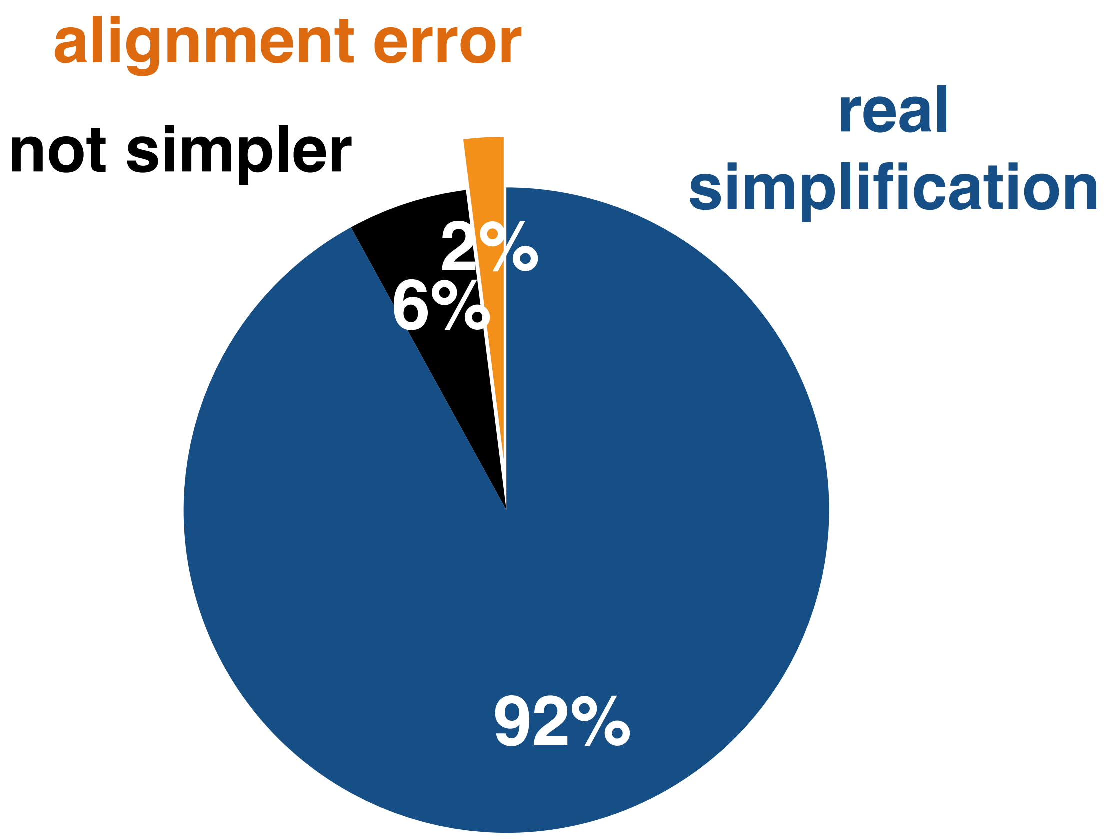
The 3,330-year-old bust of Nefertiti sits in an exhibition in the Kulturforum in Berlin, Germany, March 1, 2005.
Photo: AP/Herbert Knosowski

The ancient Egyptian Queen Nefertiti has long been at the center of a mystery.

For years, archaeologists have wondered where her tomb might be hidden. Nefertiti belonged to the family line of the famous King Tutankhamun, better known as King Tut. Indeed, some believe she was Tut's mother. While the other royals in her line are

Wei Xu, Chris Callison-Burch, Courtney Napoles. “Problems in Current Text Simplification Research: New Data Can Help” (TACL 2015)





manual inspection of aligned sentence pairs

see other analyses in the paper and EMNLP '15 talk (https://vimeo.com/150290363)

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

Lesson #2: Evaluation is difficult (yet very important).

Human Evaluation

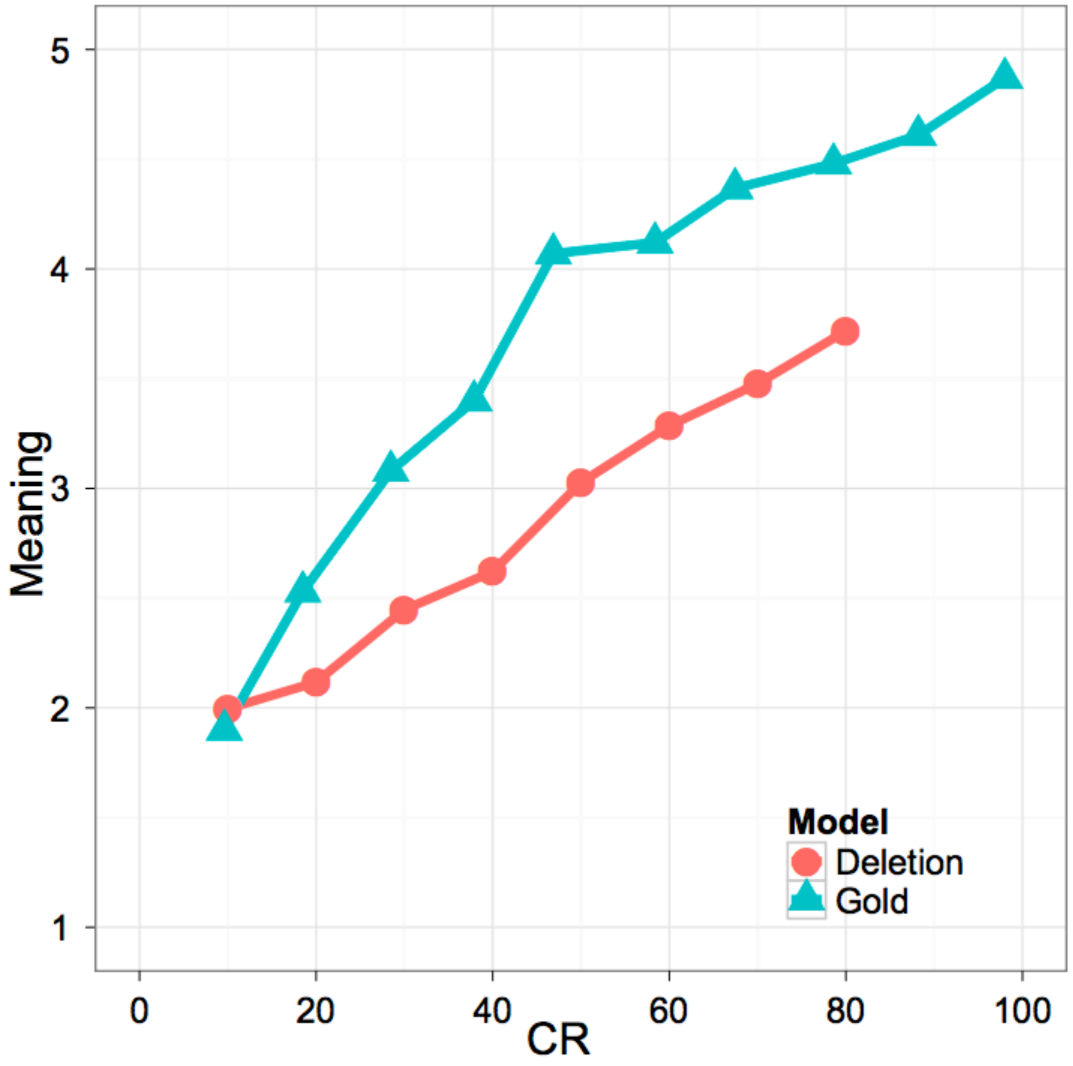
Pitfall: common practice is not necessarily good/correct

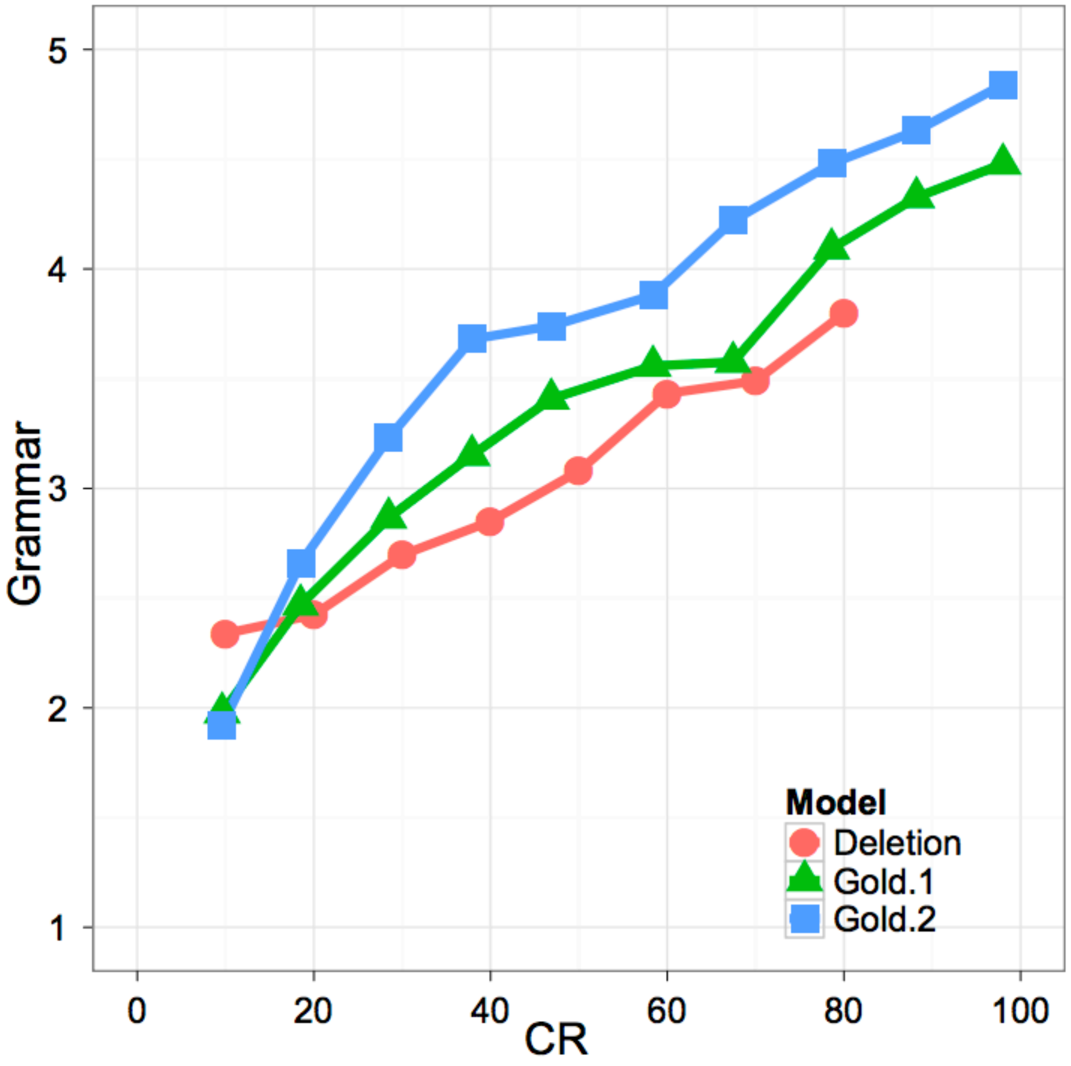
5-point Likert scale

- ● **grammaticality**
- ● **meaning preservation**

Sentence Compression

Compression rate (CR) strongly correlates wit human judgements of meaning and grammaticality





Courtney Napoles, Benjamin Van Durme, and Chris Callison-Burch. “Evaluating Sentence Compression: Pitfalls and Suggested Remedies” (T2T-G 2011)

Sentence Simplification

Pitfall: Unfairly bias towards deletion over paraphrasing

5-point Likert scale

- ● **grammaticality**
- ● **meaning preservation**
- ● **simplicity**

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, Chris Callison-Burch. "Optimizing Statistical Machine Translation for Text Simplification" (TACL 2016)

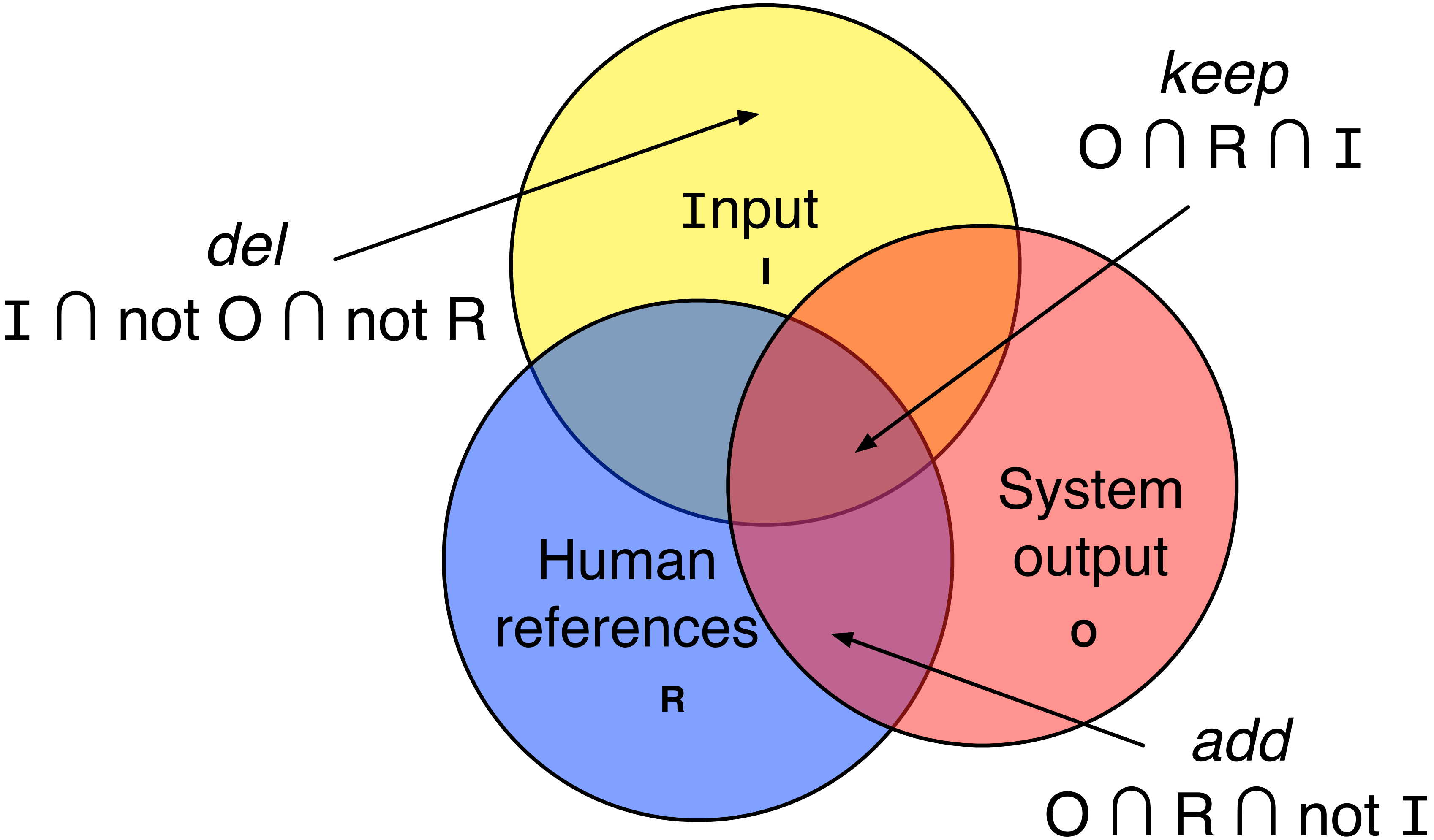
Automatic Evaluation

Pitfall: outputs with no change get high **BLEU** and high meaning and grammaticality

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, Chris Callison-Burch. "Optimizing Statistical Machine Translation for Text Simplification" (TACL 2016)

Automatic Evaluation

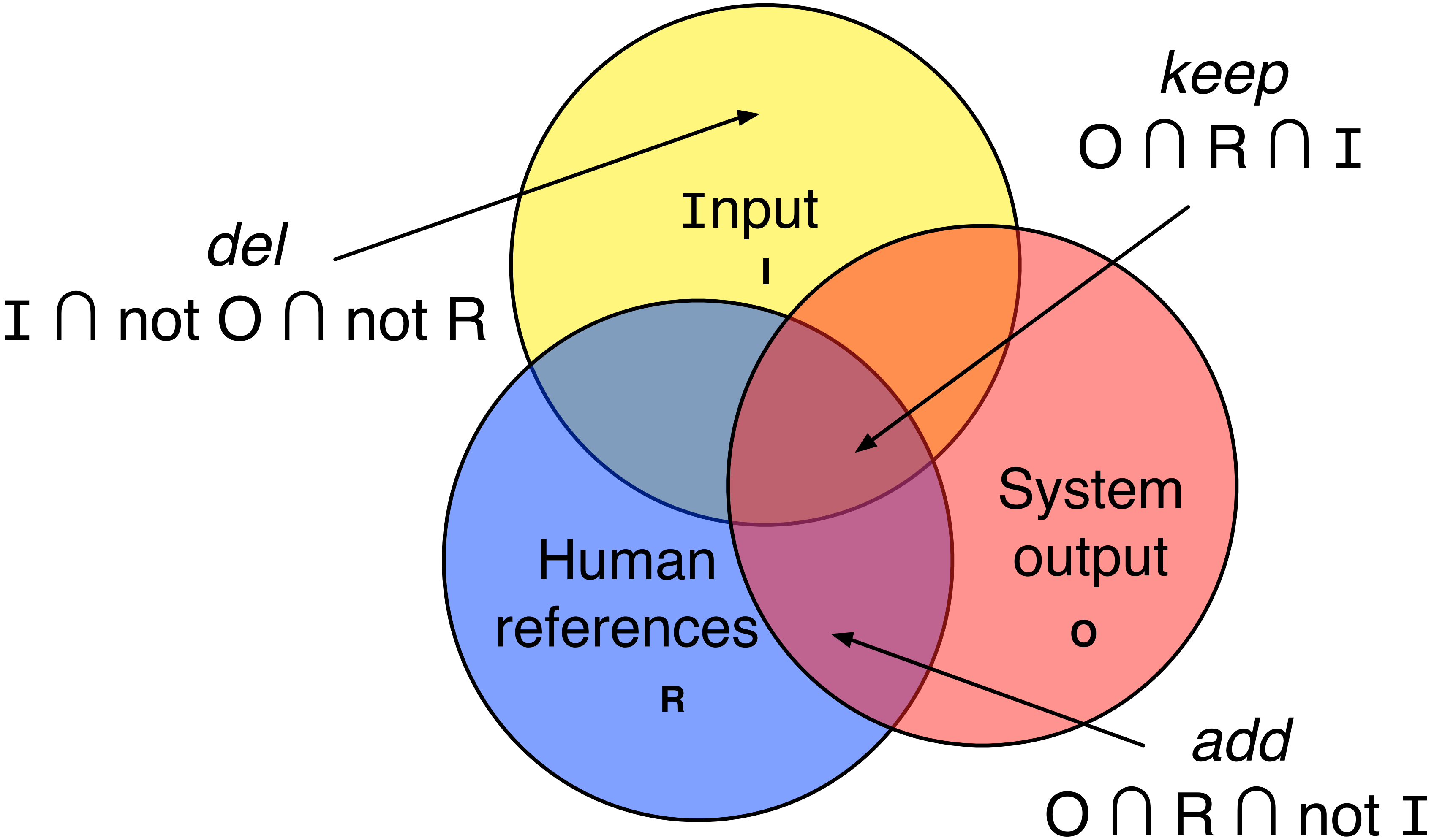
SARI tunable metric: Compare **S**ystem output **a**gainst **R**eference sentences and against the **I**nput sentence.



Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, Chris Callison-Burch. "Optimizing Statistical Machine Translation for Text Simplification" (TACL 2016)

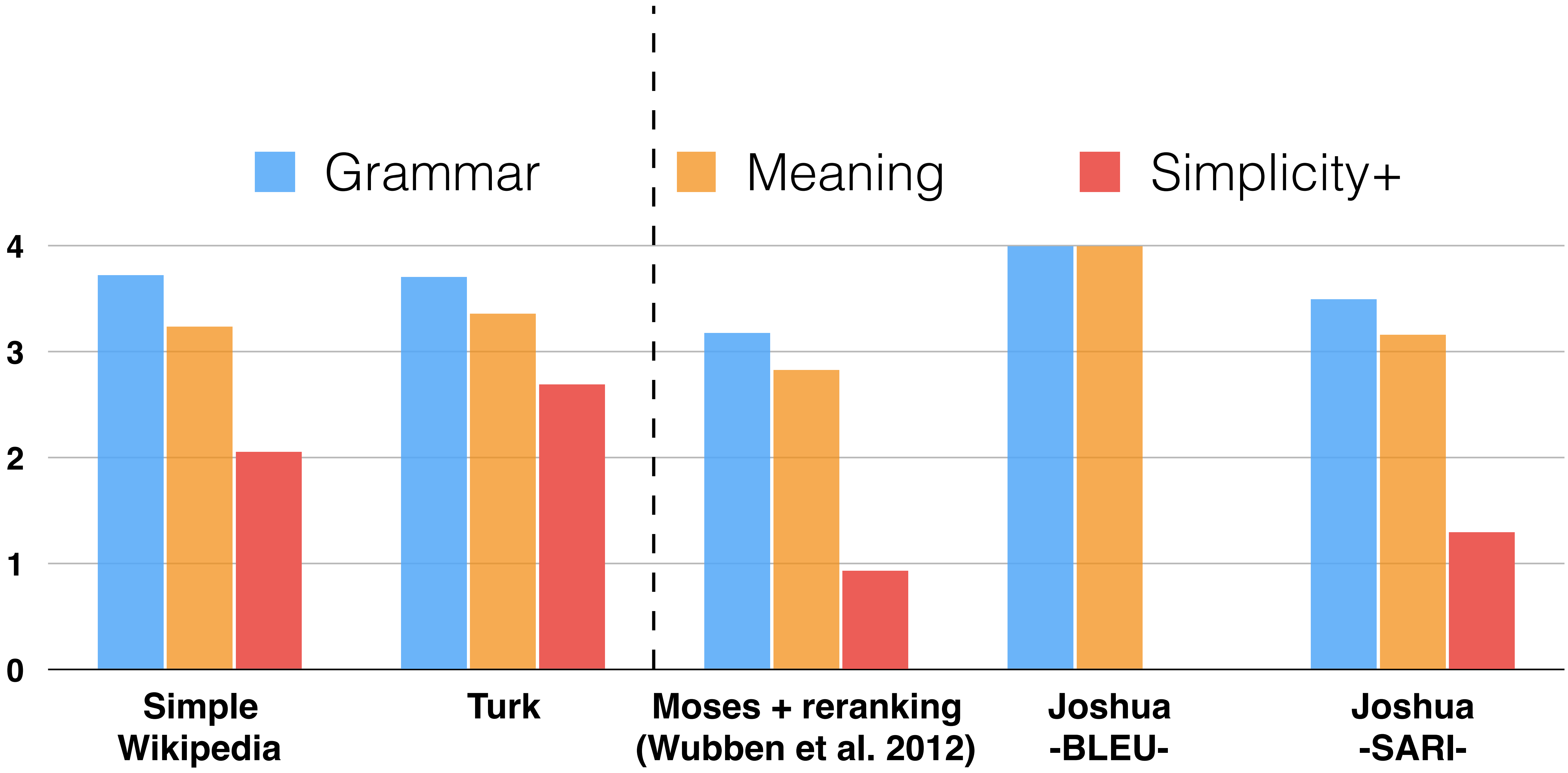
Automatic Evaluation

SARI tunable metric: Compare **S**ystem output **a**gainst **R**eference sentences and against the **I**nput sentence.



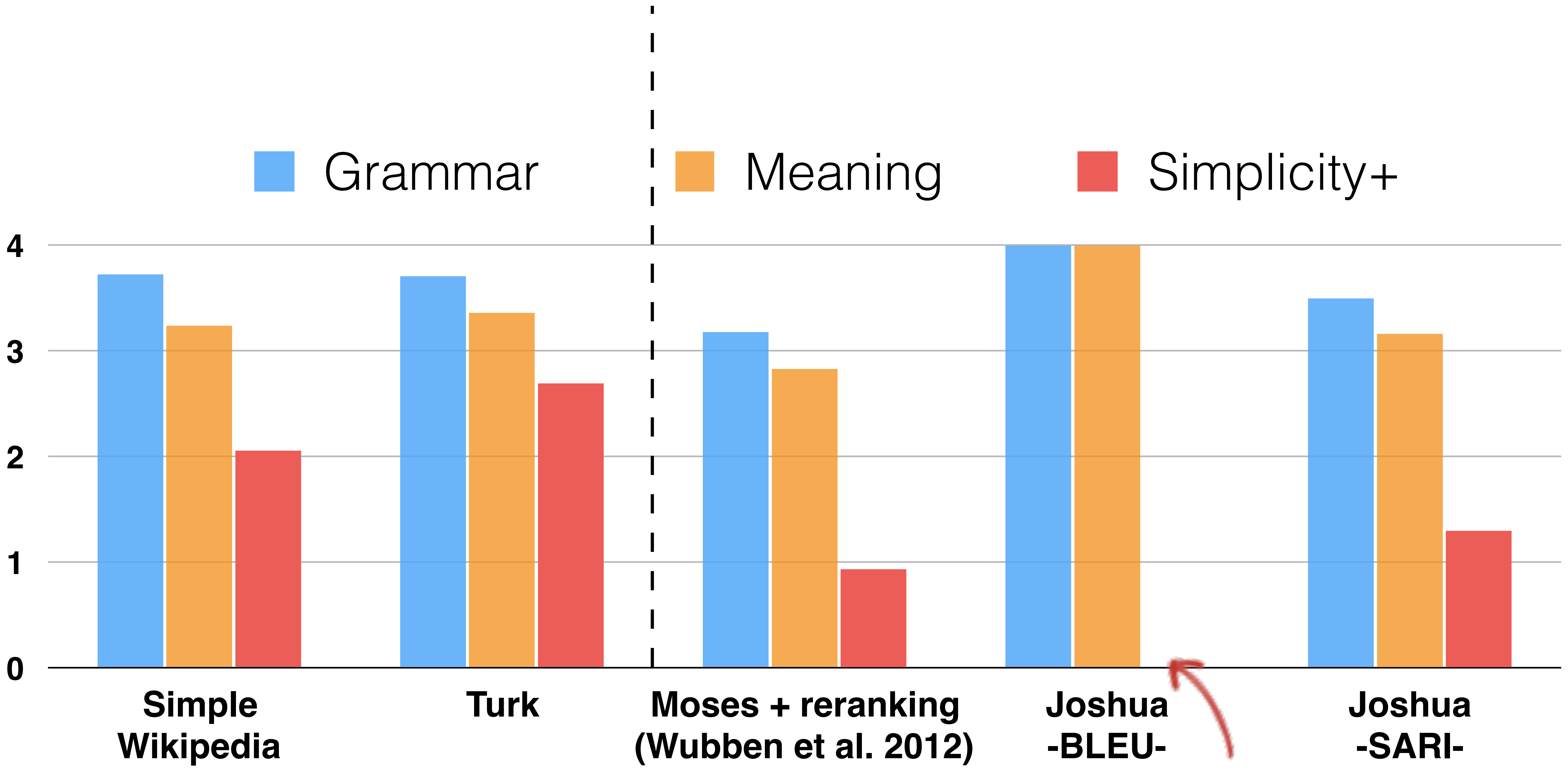
Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, Chris Callison-Burch. "Optimizing Statistical Machine Translation for Text Simplification" (TACL 2016)

Automatic Simplification



Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, Chris Callison-Burch. "Optimizing Statistical Machine Translation for Text Simplification" (TACL 2016)

Automatic Simplification

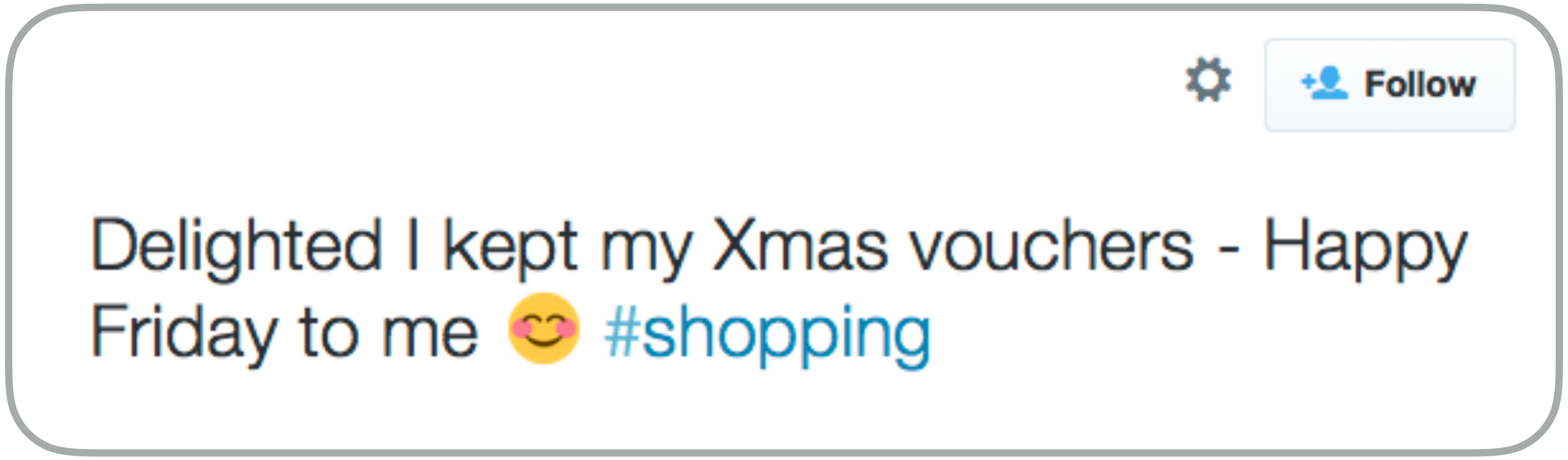


tuning towards BLEU leaves input unchanged

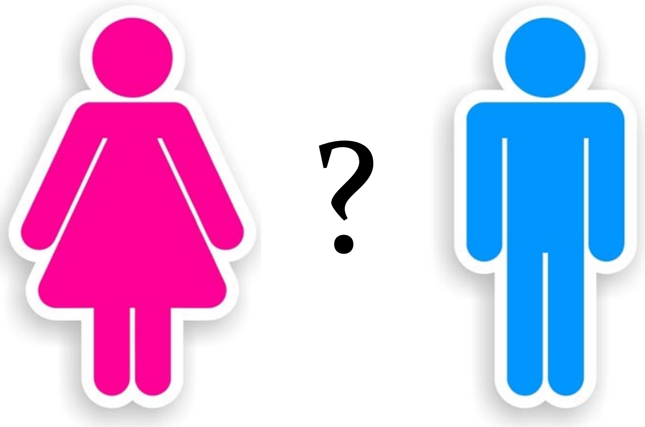
Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, Chris Callison-Burch. "Optimizing Statistical Machine Translation for Text Simplification" (TACL 2016)

Lesson #3: Linguistic styles often conflate w/ topics.

User Profiling

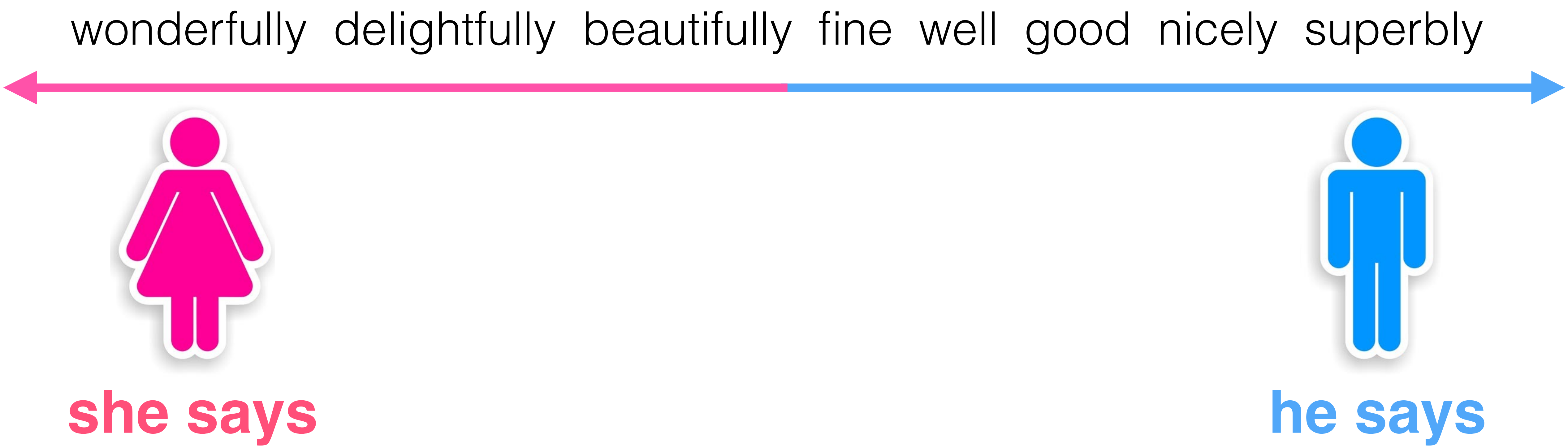


Delighted I kept my Xmas vouchers - Happy Friday to me 😊 #shopping



Source: Volkova, Van Durme, Yarowsky, Bachrach “Tutorial on Social Media Predictive Analytics” (NAACL 2015)

Paraphrase can help control topic and focus on 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” (AAAI 2016)

Style is often subtle

Good News: “Wisdom of the Crowd”
aggregated human judgements are pretty good

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

Style is often subtle

Good News: “Wisdom of the Crowd”
aggregated human judgements are pretty good

Bad News: Text-to-Text Generation in famine or masculine styles is much more difficult than Shakespearean.

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

Lesson #4: Our language is ever evolving.

Oxford Dictionaries revealed this week the earliest known usage of word "selfie" is from a 2002 online ABC forum post.

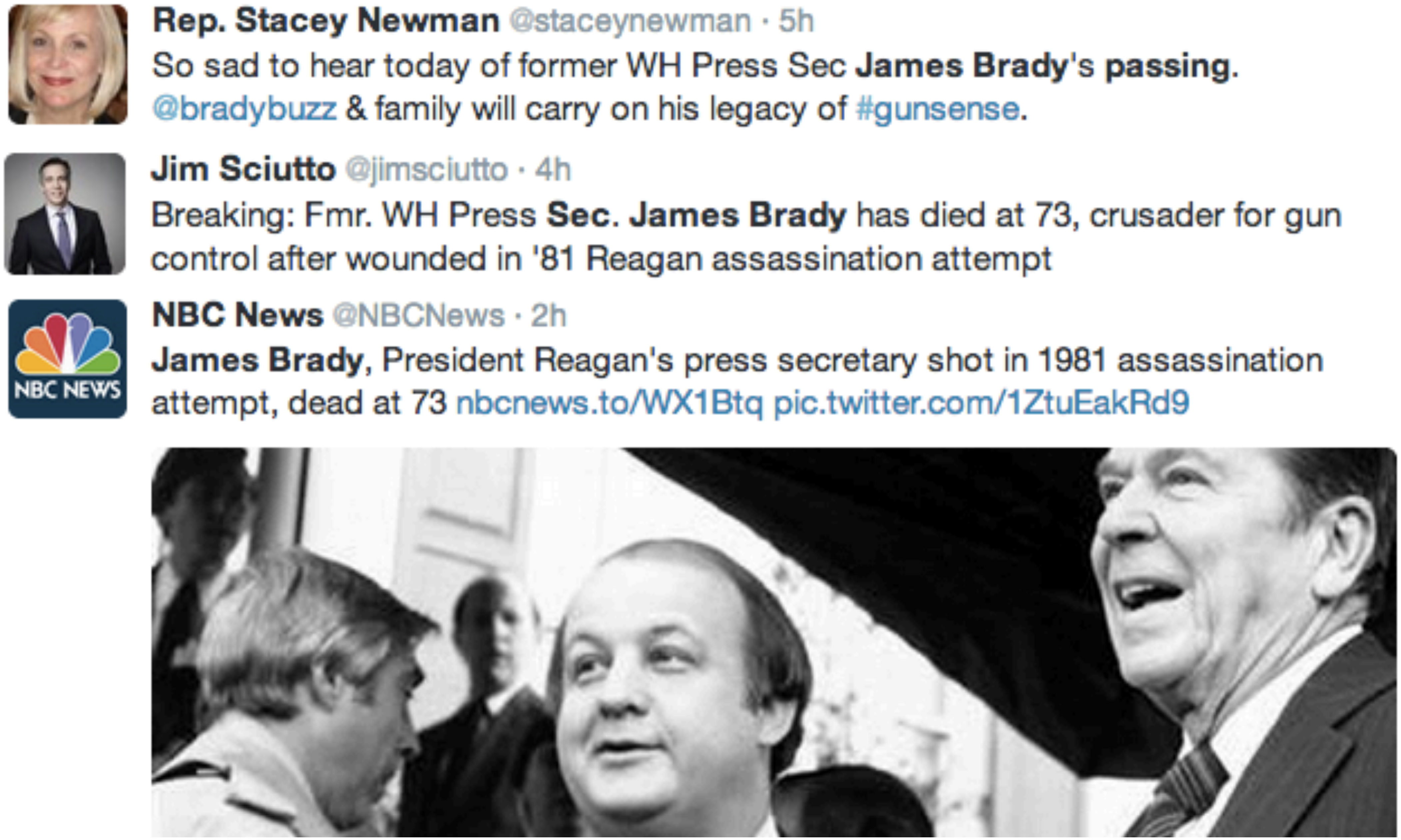
My take on Evolving Language

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)

Twitter is a powerful 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, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. "Extracting Lexically Divergent Paraphrases from Twitter" (TACL 2014)
Wei Xu, Chris Callison-Burch, Bill Dolan. "SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter" (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" (TACL 2014)
Wei Xu, Chris Callison-Burch, Bill Dolan. "SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter" (SemEval 2015)

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 …

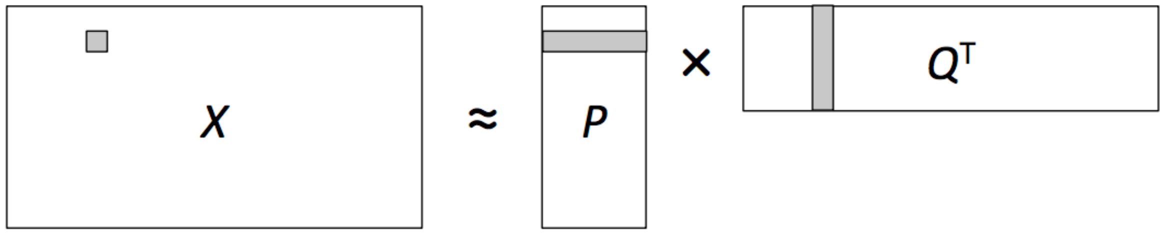
[Das & Smith 2014; Socher et al. 2011; Ling et al. 2013; Ji & Eisenstein 2013; Parikh et al. 2016; Witting & Gimpel 2017; and many others]

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

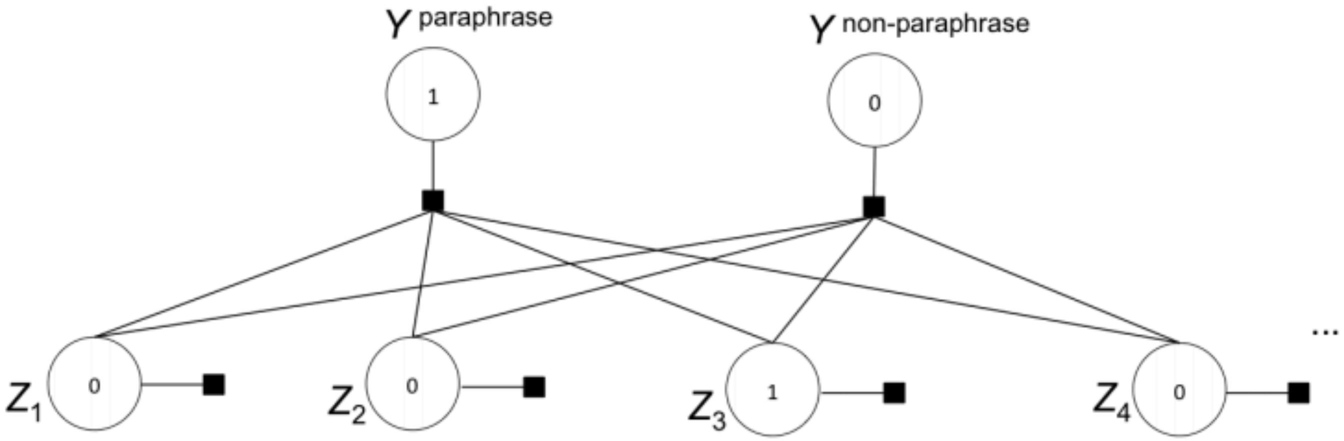
Wei Xu, Alan Ritter, Chris Callison-Burch, Bill Dolan, Yangfeng Ji. “Extracting Lexically Divergent Paraphrases from Twitter” (TACL 2014)

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

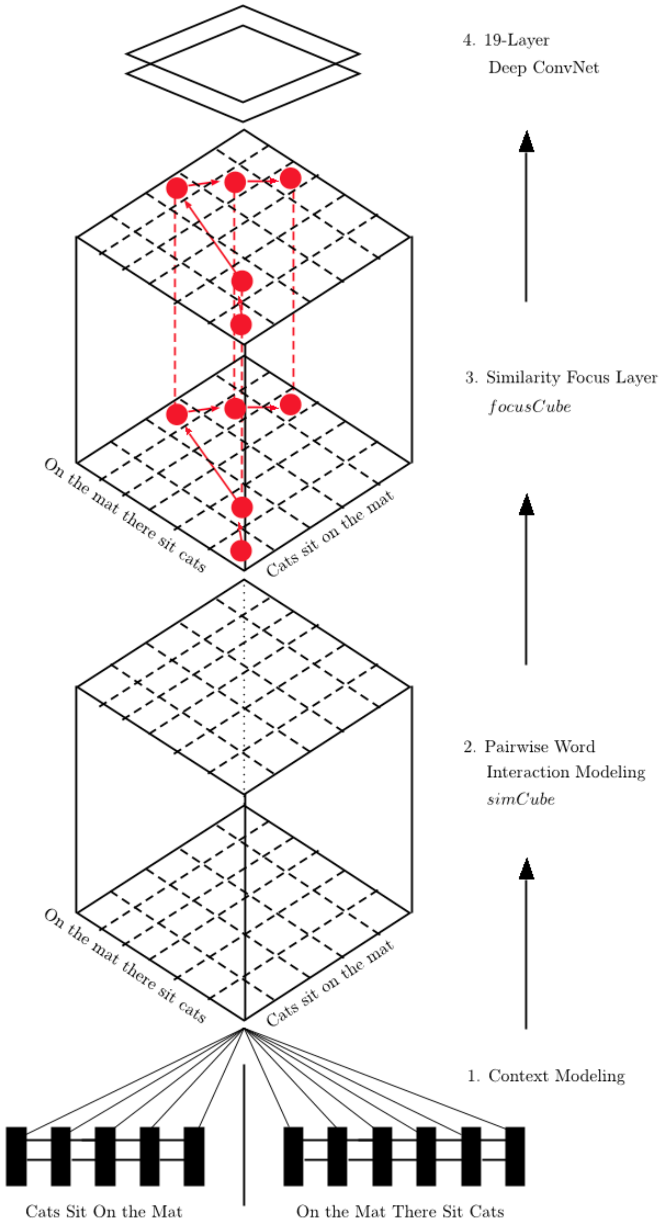
Automatic Paraphrase Identification + Word Alignment



- **LEX-OrMF** (Orthogonal Matrix Factorization) [Guo and Diab 2012]
- **DeepPairwiseWord** (Deep Neural Networks) [He et al. 2015; Ongoing Work]
- **MultiP** (Multiple Instance Learning) [Xu et al. 2014; Ongoing Work]



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


4. 19-Layer Deep ConvNet
3. Similarity Focus Layer focusCube
On the mat there sit cats
Cats sit on the mat
2. Pairwise Word Interaction Modeling simCube
On the mat there sit cats
Cats sit on the mat
1. Context Modeling
Cats Sit On the Mat
On the Mat There Sit Cats

Wuwei Lan, Siyu Qiu, Hua He, Wei Xu. “A Continuously Growing Dataset of Sentential Paraphrases” (EMNLP 2017)
Wuwei Lan, Wei Xu. “A Better Pairwise Neural Model” Ongoing Work

[Twitter Paraphrase Corpus]

51,524 sentence pairs (manually annotated)

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

Sep-10 10:30am poster

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

Timely Paraphrases

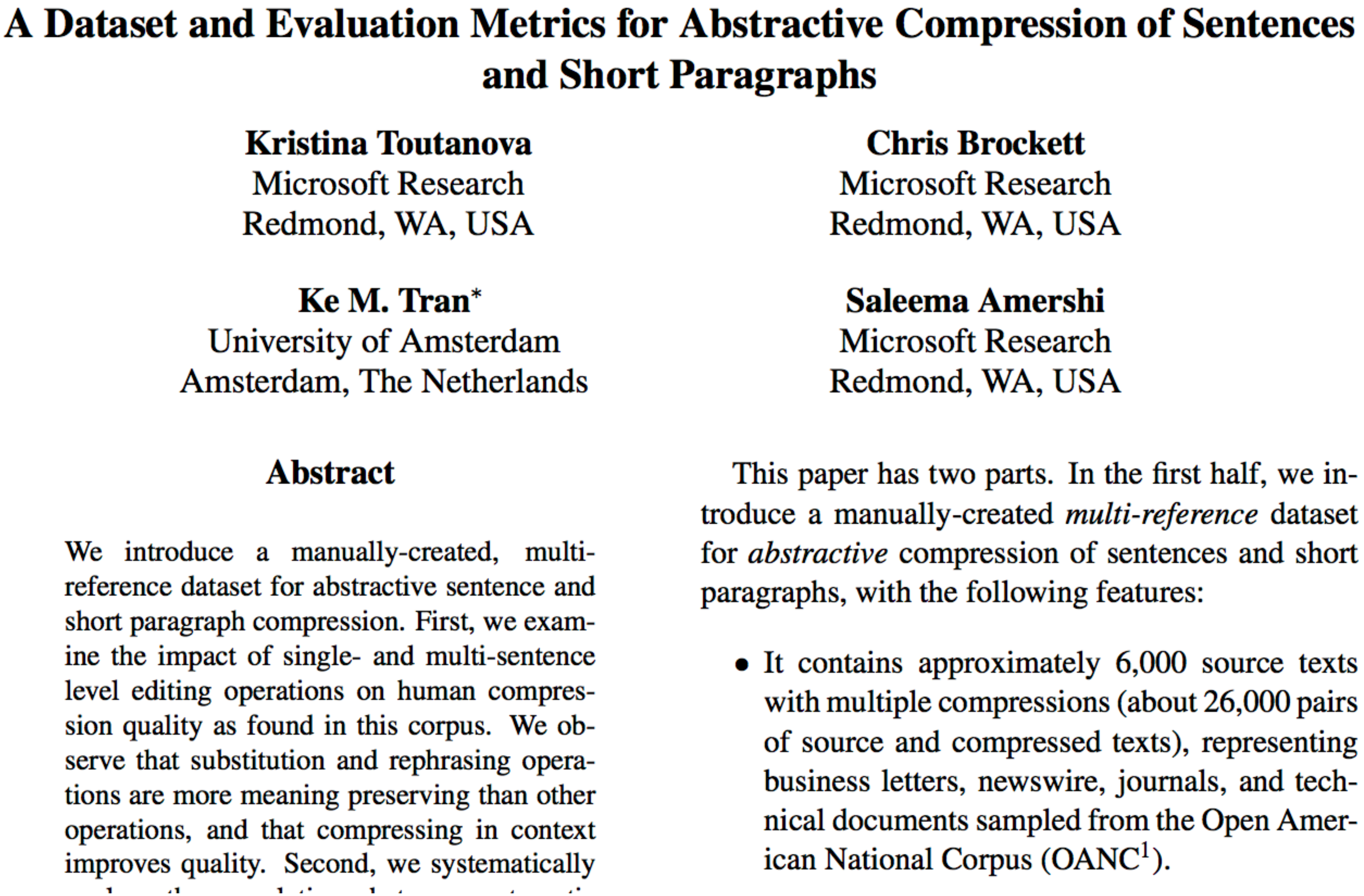
Donald Trump, DJT, Drumpf, Mr Trump, Idiot Trump, Chump, Evil Donald, #OrangeHitler, Donald @realTrump, D*nald Tr*mp, Comrade #Trump, Crooked #Trump, CryBaby Trump, Daffy Trump, Donald KKKrump, Dumb Trump, GOPTrump, Incompetent Trump, He-Who-Must-Not-Be-Named, Pres-elect Trump, President-Elect Trump, President-elect Donald J . Trump, PEOTUS Trump, Emperor Trump

Sep-10 10:30am poster

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

Lesson #5: Data! Data! Data!

Sentence Compression Dataset



A Dataset and Evaluation Metrics for Abstractive Compression of Sentences and Short Paragraphs

Kristina Toutanova
Microsoft Research
Redmond, WA, USA

Chris Brockett
Microsoft Research
Redmond, WA, USA

Ke M. Tran*
University of Amsterdam
Amsterdam, The Netherlands

Saleema Amershi
Microsoft Research
Redmond, WA, USA

Abstract

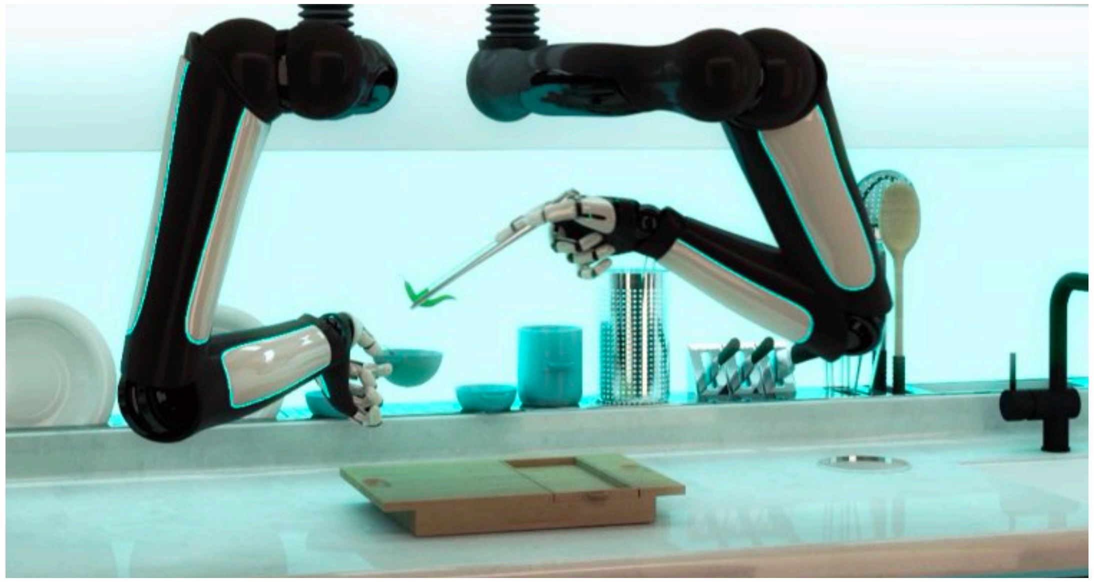
We introduce a manually-created, multi-reference dataset for abstractive sentence and short paragraph compression. First, we examine the impact of single- and multi-sentence level editing operations on human compression quality as found in this corpus. We observe that substitution and rephrasing operations are more meaning preserving than other operations, and that compressing in context improves quality. Second, we systematically

This paper has two parts. In the first half, we introduce a manually-created *multi-reference* dataset for *abstractive* compression of sentences and short paragraphs, with the following features:

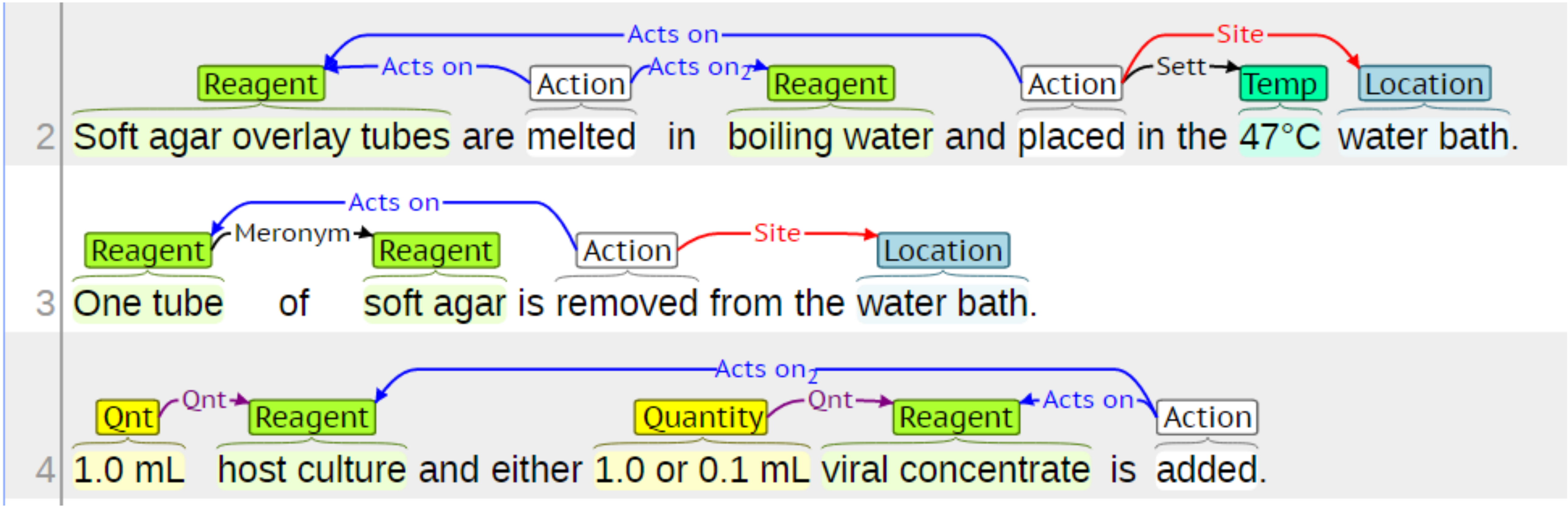
- It contains approximately 6,000 source texts with multiple compressions (about 26,000 pairs of source and compressed texts), representing business letters, newswire, journals, and technical documents sampled from the Open American National Corpus (OANC[1]).

(EMNLP 2016)

Instructional Language



Wet Lab Protocols:



2 Soft agar overlay tubes are melted in boiling water and placed in the 47°C water bath.

3 One tube of soft agar is removed from the water bath.

4 1.0 mL host culture and either 1.0 or 0.1 mL viral concentrate is added.

Chaitanya Kulkarni, Wei Xu, Alan Ritter, Raghu Machiraju
“Taking the First Essential Steps in Automating the Wet Laboratory: Annotating a Corpus of Protocols for Reproducibility” (Ongoing Work)

We Can Do It!

as long as
we have data

Thank You

thanku

thank u 4 ur time

thanking you

gramercies

thanks a lot

appreciate it

gratitude

thx

3x

tyvm

thanks

say thanks

thank you very much

thnx

thanks a ton

I can no other answer make but thanks,
And thanks, and ever thanks.

wawwww thankkkkkkkkkk you alottttttttttt!

I am grateful

Follow @cocoweixu

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