From Shakespeare to Twitter:
What are Language Styles all about?

Wei Xu
Department of Computer Science and Engineering
THE OHIO STATE UNIVERSITY
Ah... Twitter!

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?

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

I trained machine translation models ...
If you will not be turned, you will be destroyed!

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)
Paraphrasing for Style

Wei Xu¹  Alan Ritter²  William B. Dolan³  Ralph Grishman¹  Colin Cherry⁴
(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.
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)
• Se Won Jang, Jesik Min, Mark Kwon. “Writing Style Conversion using Neural Machine Translation” (Stanford CS224n 2017)

Paraphrasing ≈ Monolingual Translation ≈ T2T Generation
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.
Most fourth-graders are better readers than they were 10 years ago.
My Research

My plan

Information Extraction
My plan

Information Extraction

Reality

Paraphrase/Stylistics is exciting!
(with many pitfalls)

My Research on **Stylistics/Paraphrase**

<table>
<thead>
<tr>
<th>Term</th>
<th>Direction</th>
<th>Term</th>
<th>Direction</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>erroneous</td>
<td>→</td>
<td>correct</td>
<td></td>
<td>(Xu et al. EMNLP ’11)</td>
</tr>
<tr>
<td>writer style</td>
<td>↔</td>
<td>plain</td>
<td></td>
<td>(Xu et al. COLING ’12)</td>
</tr>
<tr>
<td>complex</td>
<td>→</td>
<td>simple</td>
<td></td>
<td>(Xu et al. TACL ’15)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Xu et al. TACL ’16)</td>
</tr>
<tr>
<td>noisy</td>
<td>↔</td>
<td>standard</td>
<td></td>
<td>(Xu et al. BUCC ’13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Xu et al. TACL ’14)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Xu et al. SemEval ’15)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Lan, Qiu, He, Xu EMNLP ’17)</td>
</tr>
<tr>
<td>feminine</td>
<td>↔</td>
<td>masculine</td>
<td></td>
<td>(Preotiu, Xu, Ungar AAAI ’16)</td>
</tr>
</tbody>
</table>
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.
Text Simplification Dataset

**Pitfall:** Simple Wikipedia is not simple.
We proposed a new high-quality dataset of news articles simplified by professional editors.
Wikipedia

- Alignment error: 17%
- Real simplification: 50%
- Not simpler: 33%

Newsela

- Alignment error: 2%
- Real simplification: 92%
- Not simpler: 6%

Manual inspection of aligned sentence pairs.

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

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 with 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](image)

- 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.
**Automatic Evaluation**

**SARI tunable metric:** Compare System output against Reference sentences and against the Input sentence.

*del*  
$I \cap \text{not } O \cap \text{not } R$

*keep*  
$O \cap R \cap I$

*add*  
$O \cap R \cap \text{not } I$

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 **System output** against **Reference** sentences and against the **Input** sentence.

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

This graph compares different methods for automatic text simplification. The x-axis represents various techniques: Simple Wikipedia, Turk, Moses + reranking (Wubben et al. 2012), Joshua -BLEU-, and Joshua -SARI-. The y-axis shows the performance in terms of grammar, meaning, and simplicity. The graph indicates that Simple Wikipedia and Turk consistently perform well, with Moses + reranking showing a notable drop in meaning and simplicity. Joshua with BLEU and SARI metrics also exhibit strong performance, especially in terms of grammar and meaning.

---

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Grammar</th>
<th>Meaning</th>
<th>Simplicity+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Wikipedia</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Turk</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Moses + reranking (Wubben et al. 2012)</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Joshua -BLEU-</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Joshua -SARI-</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

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)

Style is often subtle

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

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

<table>
<thead>
<tr>
<th>selfie</th>
<th>word</th>
<th>photo</th>
</tr>
</thead>
<tbody>
<tr>
<td>gets the boot from</td>
<td>phrase</td>
<td>has been sacked by</td>
</tr>
</tbody>
</table>

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

Jeremy Corbyn is a closest Brexiteer.

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)
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)
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]
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]

\[
P(z_i, y_i | w_i; \theta) = \prod_{j=1}^{m} \exp(\theta \cdot f(z_j, w_j)) \times \sigma(z_i, y_i)
\]

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)

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*nard 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
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

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

Chris Brockett  
Microsoft Research  
Redmond, WA, USA

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 (OANC1).
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.
We Can Do It!
as long as we have data
Thank You

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

wawwww thankkkkkkkkkkkkkk you alottttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt