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
“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)

\[
\begin{align*}
\text{meaning} & \quad = \quad \text{invariant of paraphrases} \\
\text{text} & \quad = \quad \text{‘virtual paraphrasing’} \\
\text{paraphrases} & \quad = \quad \text{synonymous linguistic expressions}
\end{align*}
\]
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

  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?

---

... 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

Worries over the health of King Bhumibol Adulyadej are shaking Thailand

Web is a Gold Mine

Web is a Gold Mine

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

- **Non-Paraphrase**
  - **Career Synchronicity** @careersync_now · 12 Oct 2016
  - Fears for King’s Health Shake Thailand ift.tt/2d7frGd

- **Non-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
  - Retweets: Likes:

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]
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” in 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)

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
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 official

UVA administrator
U-Va. dean

Noisy Text Normalization

Translate

Hostes is going outta biz.

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” 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

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)
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)
We speak in style!

wonderfully  delightfully  beautifully  fine  well  good  nicely  superbly

she says

he says

(also age & income)


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

- erroneous $\rightarrow$ correct
  - (Xu et al. EMNLP ’11)
- writer style $\leftrightarrow$ plain
  - (Xu et al. COLING ’12)
- complex $\rightarrow$ simple
  - (Xu et al. TACL ’15)
  - (Xu et al. TACL ’16)
- noisy $\rightarrow$ standard
  - (Xu et al. BUCC ’13)
  - (Xu et al. TACL ’14)
  - (Xu et al. SemEval ’15)
- feminine $\leftrightarrow$ masculine
  - (Preotiuc, Xu, Ungar AAAI ’16)

and more language variations (future work) …
After all, we want a robot that can talk and take selfies ...
Thank You

thanking you

appreciate it

thank you very much

thnx

say thanks

thanks

thank you very much

wawww thankkkkkkkkkkkk you alotttttttttttttt!

thanks a lot

I am grateful

thank u 4 ur time

thx

3x

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