Paraphrases in Twitter

Wei Xu
Computer and Information Science
University of Pennsylvania
Paraphrase
... the forced resignation of the CEO of Boeing, Harry Stonecipher, for ...

Boeing Co. Monday said it fired Chief Executive Harry Stonecipher, after ...
Application

Information Extraction

end_job (Harry Stonecipher, Boeing)
title (Harry Stonecipher, CEO)

… the forced resignation of the CEO of Boeing, Harry Stonecipher, for …

Boeing Co. Monday said it fired Chief Executive Harry Stonecipher, after …

Who is the CEO stepping down from Boeing?
They are culturally akin to the coastal peoples of Papua New Guinea.

Their culture is like that of the coastal peoples of Papua New Guinea.

Application

Stylistic Rewriting

Palpatine:
If you will not be turned, you will be destroyed!

Luke:
Father, please! Help me!

Father, I pray you! Help me!

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

Numerous publications for paraphrase identification, extraction, generation and various applications.

But, primarily for formal language and well-edited text.
Previous Work

only a few hundreds news agencies report big events in formal language

(Dolan, Quirk and Brockett, 2004; Dolan and Brockett, 2005; Brockett and Dolan, 2005)
Twitter as a new 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
Twitter as a powerful resource

thousands of users
talk about both big and micro events
in formal, informal, erroneous languages
Enables new applications

Human-computer Interaction

who want to get a beer?

wants to get a beer?
someone to get a beer?
who wants to go get a beer?
trying to get a beer?
who wants to buy a beer?
who wants to get a drink?
who else wants to get a beer?

… (21 different ways)

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

Noisy Text Normalization

- Oscar-nom’d doc
- don’t wait for

Oscar-nominated documentary
don’t wait for

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)
Enables new applications

Language Education

Aaaaaaaaaaand stephen curry is on fire

What a incredible performance from Stephen Curry
Enables new applications

Sentiment Analysis  😊 or 😞  ?

This nets vs bulls game is great
This Nets vs Bulls game is nuts
Wowsers to this nets bulls game
This Nets and Bulls game is bananas
this Nets vs Bulls game is too live
This Nets and Bulls game is a good game
This netsbulls game is too good
This NetsBulls series is intense
Trilogy of Paraphrase

# 1
Trilogy of Paraphrase

#1 identify parallel sentences

Mancini has been sacked by Manchester City
Mancini gets the boot from Man City

Yes!

WORLD OF JENKS IS ON AT 11
World of Jenks is my favorite show on tv

No!
Design a model
Train it by data
Design a model

At-least-one-anchor Assumption

two sentences about the same topic are paraphrases if and only if they contain at least one word pair that is a paraphrase anchor

That boy Brook Lopze with a deep 3
brook lopez hit a 3

Yes!
A challenge

not every word pair of similar meaning indicates sentence-level paraphrase

<table>
<thead>
<tr>
<th>Iron Man 3 was brilliant fun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iron Man 3 tonight see what this is like</td>
</tr>
</tbody>
</table>

Solution:

a discriminative model using features at word-level
Multi-instance Learning Paraphrase Model

Manti bout to be the **next** Junior Seau

Teo is the little **new** Junior Seau

---

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan and Yangfeng Ji. “Gathering and Generating Paraphrases from Twitter with Application to Normalization” In Transactions of the Association for Computational Linguistics (2014)
[Mini Tutorial]  
Multi-instance Learning

Instead of labels on each individual instance, the learner only observes labels on bags of instances.

A bag is labeled negative, if all the examples in it are negative

A bag is labeled positive, if there is at least one positive example

(Dietterich et al., 1997)
Multi-instance Learning

Latent Variable Model

A bag is labeled positive, if there is at least one positive example

(Riedel et al. 2010; Hoffmann et al., 2011; Surdeanu et al. 2012)

Wei Xu, Raphael Hoffmann, Le Zhao, Ralph Grishman. “Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction”

In ACL (2013)
Multi-instance Learning

Latent Variable Model

A bag is labeled negative, if all the examples in it are negative.

(Riedel et al. 2010; Hoffmann et al., 2011; Surdeanu et al. 2012)

Wei Xu, Raphael Hoffmann, Le Zhao, Ralph Grishman. “Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction”
In ACL (2013)
[Recap] Multi-instance Learning Paraphrase Model

\[ \begin{align*}
Z_1 & \quad \gamma_{\text{paraphrase}} \quad 1 \\
Z_2 & \quad 0 \\
Z_3 & \quad 1 \\
Z_4 & \quad 0
\end{align*} \]

- \text{manti | teo}
- \text{be | is}
- \text{next | new}
- \text{manti | little}

Features:
- \text{diff_word}
- \text{same_pos_nn}
- \text{both_sig}
- \text{...}
- \text{same_stem}
- \text{same_pos_be}
- \text{not_both_sig}
- \text{...}
- \text{diff_word}
- \text{same_pos_jj}
- \text{both_sig}
- \text{...}
- \text{diff_word}
- \text{diff_pos_nn}
- \text{diff_pos_jj}
- \text{not_both_sig}
- \text{...}
Joint word-sentence inference

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

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

deterministic OR

\[
\sigma(z_i, y_i) = \begin{cases} 
1 & \text{if } y_i = true \land \exists j : z_j = 1 \\
1 & \text{if } y_i = false \land \forall j : z_j = 0 \\
0 & \text{otherwise}
\end{cases}
\]
Learning algorithm

Objective

\[ \theta^* = \arg \max_{\theta} P(y|w; \theta) \]
\[ = \arg \max_{\theta} \prod_i \sum_{z_i} P(z_i, y_i|w_i; \theta) \]

all sentence pairs in training set

Update

\[ \frac{\partial \log P(y|w; \theta)}{\partial \theta} = E_{P(z|w,y;\theta)} \left( \sum_i f(z_i, w_i) \right) \]
\[ - E_{P(z,y|w;\theta)} \left( \sum_i f(z_i, w_i) \right) \]
\[ \approx \sum_i f(z_i^*, w_i) - \sum_i f(z'_i, w_i) \]
[Recap] Task

#1 identify parallel sentences

Mancini has been sacked by Manchester City

Mancini gets the boot from Man City

Yes!

WORLD OF JENKS IS ON AT 11

World of Jenks is my favorite show on tv

No!
[Recap] Model

observation

<table>
<thead>
<tr>
<th>That boy Brook Lopez with a deep 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>brook lopez hit a 3</td>
</tr>
</tbody>
</table>

Yes!

Multi-instance Learning Paraphrase Model (MultiP)

- big data stream
- principled latent variable model
- flexible extensible framework
- no word-level annotation needed
Training data
Training data

Crowdsourcing

(Courtesy: The Sheep Market)
Training data

Crowdsourcing

Sentence: **Borussia Dortmund advanced to the final**

Select ALL sentences that have similar meaning from below:

- Borussia Dortmund has clinched their Champions League final spot
- Real Madrid efforts are not enough as Cinderella Borussia Dortmund advances to the Champions League Final
- But it’s Borussia Dortmund whose heading to Wembley Park
- Congratulations Borussia Dortmund is going to Wembley
A problem

only 8% sentence pairs about the same topic have similar meaning

hurt both quantity and quality

non-experts lower their bars
Sentence Selection

8% → 16%

Salience(s) = \sum_{w_i \in s} \frac{P(w_i)}{|w_i| w_i \in s}

- U.S.
- Facebook
- Dwight Howard
- GWB
- Netflix
- Ronaldo
- Dortmund
- Momma Dee
- Morning
- Huck
- Klay
- Milwaukee
- Harvick
- Jeff Green
- Ryu
- The Clippers
- Candice
- Candice
- Robert Woods
- Amber
- Reggie Miller

[Diagram showing the comparison of filtered and random percentage of positive judgements for various trending topics.]

(filtered vs. random)
Topic Selection

Multi-Armed Bandits

16% → 34%

\[
\begin{align*}
\max & \quad \sum_{j \in \{i \mid r_i(t_0) > 0\}} \hat{\mu}_i(t_0) r_i(t_1) \\
\text{s.t.} & \quad \sum_{j \in \{i \mid r_i(t_1) > 0\}} r_i(t_0) \leq (1 - \epsilon) B, \forall i: 0 \leq r_i(t_1) \leq l - r_i(t_0).
\end{align*}
\]
Training data

18,762 sentence pairs labeled cost only $200
1/3 paraphrase, 2/3 non-paraphrase (very balanced)

including very broad-range paraphrases: synonyms, misspellings, slang, acronyms and colloquialisms

Very diverse!
Experiment
Experiment

![Bar chart showing precision and recall for paraphrase identification models.](chart)

- **Das & Smith (2009)**: Precision 62.9, Recall 63.2
- **Guo & Diab (2012)**: Precision 52.5, Recall 65.5
- **Ji & Eisenstein (2013)**: Precision 66.4, Recall 62.8
- **Our Model**: Precision 72.2, Recall 72.6
- **Human**: Precision 75.2, Recall 90.8

The human performance shows state-of-the-art results on paraphrase identification.
Semantic Evaluation 2015

based on our Twitter paraphrase dataset
19 + 1 teams participated
some used neural networks

paraphrase identification (0 or 1) rank 1 🏆
semantic similarity (0 ~ 1) rank 4
Trilogy of Paraphrase

# 2 & 3
#2 extract lexical/phrasal paraphrase

Mancini has been sacked by Manchester City

align

Mancini gets the boot from Man City
#2 extract lexical/phrasal paraphrase

<table>
<thead>
<tr>
<th>has been sacked by</th>
<th>gets the boot from</th>
</tr>
</thead>
<tbody>
<tr>
<td>manchester city</td>
<td>man city</td>
</tr>
<tr>
<td>4</td>
<td>for</td>
</tr>
<tr>
<td>4</td>
<td>four</td>
</tr>
<tr>
<td>outta</td>
<td>out of</td>
</tr>
<tr>
<td>hostes</td>
<td>hostess</td>
</tr>
</tbody>
</table>
Trilogy of Paraphrase

#3  text-to-text generation

Hostes is going outta biz.

translate

Hostess is going out of business.
Statistical Machine Translation

<table>
<thead>
<tr>
<th>Feature</th>
<th>Bilingual</th>
<th>Monolingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>studied</td>
<td>a lot</td>
<td>getting more</td>
</tr>
<tr>
<td>sensitive to error</td>
<td>less</td>
<td>more</td>
</tr>
<tr>
<td>naturally available parallel text</td>
<td>more</td>
<td>less</td>
</tr>
<tr>
<td>objective</td>
<td>simple</td>
<td>flavored</td>
</tr>
<tr>
<td>has standard evaluation</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>
Text-to-text Generation

- historic ↔ modern
- noisy → standard
- complex → simple
- prose → sonnet
Wandering through rows of stalls examining workhorses and prize hogs may seem to ... have been a strange way for a scientist to spend an afternoon, but there was a certain logic to it.

The Ideal
The Ideal

You've suffered a contusion to the soft tissue below the fourth thoracic vertebra, exacerbating the proximal sternum.

Translation: "You have a bruised rib."
Collaborators

Chris Callison-Burch  UPenn
Bill Dolan  MSR
Alan Ritter  UW / OSU
Ralph Grishman  NYU
Raphael Hoffmann  UW / AI2 Incubator
Joel Tetreault  ETS / Yahoo!
Le Zhao  CMU / Google
Colin Cherry  NRC
Yangfeng Ji  GaTech
Thank you

thanking you

thx
tyvm
say thanks
thnx

wawwww thankkkkkkkkkkkkkk you alotttttttttttttt!

thanks a lot

am grateful

thank you very much

appreciate it

thank u 4 ur time

gratitude
3x

more details on my homepage: http://www.cis.upenn.edu/~xwe/