Problems in Current Text Simplification Research

Wei Xu
Chris Callison-Burch
Courtney Napoles

UPenn
UPenn
JHU
What is Text Simplification

What is Text Simplification

INPUT

Applesauce is a puree made of apples.
What is **Text Simplification**

**INPUT**  
*Applesauce is a puree made of apples.*

**OUT-1**  
*Applesauce is a soft paste.*
What is Text Simplification

INPUT  Applesauce is a puree made of apples.
OUT-1  Applesauce is a soft paste.
OUT-2  Applesauce is a paste. It is made of apples.
What is Text Simplification

- for children, disabled, non-native speakers ...
- for other NLP tasks (MT, summarization ...)
What is Text Simplification

paraphrasing

INPUT  Applesauce is a puree made of apples.
OUT-1  Applesauce is a soft paste.
OUT-2  Applesauce is a paste. It is made of apples.

• for children, disabled, non-native speakers ...
• for other NLP tasks (MT, summarization ...)
What is **Text Simplification**

- **paraphrasing**
- **deletion**

**INPUT**  
*Applesauce is a puree made of apples.*

**OUT-1**  
*Applesauce is a soft paste.*

**OUT-2**  
*Applesauce is a paste. It is made of apples.*

- for children, disabled, non-native speakers …
- for other NLP tasks (MT, summarization …)

What is **Text Simplification**

- paraphrasing
- deletion
- splitting

**INPUT**  
*Applesauce is a puree made of apples.*

**OUT-1**  
*Applesauce is a soft paste.*

**OUT-2**  
*Applesauce is a paste. It is made of apples.*

- for children, disabled, non-native speakers …
- for other NLP tasks (MT, summarization …)

What is Text Simplification

- paraphrasing
- deletion
- splitting

INPUT  
**Applesauce is a puree made of apples.**

OUT-1  
**Applesauce is a soft paste.**

OUT-2  
**Applesauce is a paste. It is made of apples.**

- for children, disabled, non-native speakers …
- for other NLP tasks (MT, summarization …)

Goal of **Text Simplification**

**INPUT**  
Applesauce is a puree made of apples.

**OUT-1**  
Applesauce is a soft paste.

**OUT-2**  
Applesauce is a paste. It is made of apples.
Goal of Text Simplification

grammaticality

INPUT  
Applesauce is a puree made of apples.

OUT-1  
Applesauce is a soft paste.

OUT-2  
Applesauce is a paste. It is made of apples.
Goal of Text Simplification

grammaticality  meaning preservation

INPUT  Applesauce is a puree made of apples.
OUT-1  Applesauce is a soft paste.
OUT-2  Applesauce is a paste. It is made of apples.
Goal of Text Simplification

- grammaticality
- meaning preservation
- simplicity

INPUT  
Applesauce is a puree made of apples.

OUT-1  
Applesauce is a soft paste.

OUT-2  
Applesauce is a paste. It is made of apples.
**Goal of Text Simplification**

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

**INPUT**  
*Applesauce is a puree made of apples.*

**OUT-1**  
*Applesauce is a soft paste.*  

**OUT-2**  
*Applesauce is a paste. It is made of apples.*

**Human Evaluation**

- **grammaticality:**
  - OUT-1: 5
  - OUT-2: 5

- **meaning preservation:**
  - OUT-1: 4
  - OUT-2: 5

- **simplicity:**
  - OUT-1: 5
  - OUT-2: 4
Goal of **Text Simplification**

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

**INPUT**

*Applesauce is a puree made of apples.*

**OUT-1**

*Applesauce is a soft paste.*

**OUT-2**

*Applesauce is a paste. It is made of apples.*

**Human Evaluation**

5 4 5

5 5 4

(no reliable automatic evaluation yet)

---

Brief History of Sentence Simplification

Brief History of Sentence Simplification

rule-based

1997 Chandrasekar & Srinivas
1999 Dras (PhD thesis)
2000 Carroll, Minnen, Pearce, Canning, Devlin
2002 Canning (PhD thesis)
2004 Siddharthan (PhD thesis)

Brief History of Sentence Simplification

1997 Chandrasekar & Srinivas
1999 Dras (PhD thesis)
2000 Carroll, Minnen, Pearce, Canning, Devlin
2002 Canning (PhD thesis)
2004 Siddharthan (PhD thesis)
2010 Zhu, Bernhard, Gurevych

rule-based
Parallel Wikipedia Corpus

Brief History of Sentence Simplification

1997  Chandrasekar & Srinivas
1999  Dras (PhD thesis)
2000  Carroll, Minnen, Pearce, Canning, Devlin
2002  Canning (PhD thesis)
2004  Siddharthan (PhD thesis)
2010  Zhu, Bernhard, Gurevych

Brief History of Sentence Simplification

1997  Chandrasekar & Srinivas
1999  Dras (PhD thesis)
2000  Carroll, Minnen, Pearce, Canning, Devlin
2002  Canning (PhD thesis)
2004  Siddharthan (PhD thesis)
2010  Zhu, Bernhard, Gurevych
2011  Woodsend & Lapata
2011  Coster & Kauchak
2012  Wubben, van den Bosch, Krahmer
2014  Narayan & Gardent
2014  Siddharthan (Survey)
2014  Angrosh, Nomoto, Siddharthan
2014  Narayan (PhD thesis)

Brief History of Sentence Simplification

1997  Chandrasekar & Srinivas
1999  Dras (PhD thesis)
2000  Carroll, Minnen, Pearce, Canning, Devlin
2002  Canning (PhD thesis)
2004  Siddharthan (PhD thesis)
2010  Zhu, Bernhard, Gurevych
2011  Woodsend & Lapata
2011  Coster & Kauchak
2012  Wubben, van den Bosch, Krahmer
2014  Narayan & Gardent
2014  Siddharthan (Survey)
2014  Angrosh, Nomoto, Siddharthan
2014  Narayan (PhD thesis)
Now  Xu, Callison-Burch, Napoles (Opinion)

Problems in Simplification Research

- State-of-the-art evaluation is suboptimal. But we have been doing this in the past 5 years*.

- Simple Wikipedia data dominated in the past 5 years. But its quality was taken for granted. It limits the scope of research.

* (Angrosh et al. 2014) tried comprehension quiz

Why this is important?

- better understanding
- better review
- more diverse research
- better data and evaluation
- better model
Why this is important?

- better understanding
- better review
- more diverse research
- better data and evaluation
- better model

Why this is important?

- better understanding
- better review
- more diverse research
- better data and evaluation
- better model

“Recently, there have been several attempts at addressing the text simplification task as a monolingual translation problem … However, they did not try to seek reasons for the success or the failure of their systems.”

—— Štajner, Béchara, Saggion (2015)
“Recently, there have been several attempts at addressing the text simplification task as a monolingual translation problem … However, they did not try to seek reasons for the success or the failure of their systems.”

—— Štajner, Béchara, Saggion (2015)

**WHY DID THIS HAPPEN?**

1. “state-of-the-art” competition
2. not easy to do

Opinion #1

Current evaluation doesn’t tell us what’s going on.
System Comparability

sub-systems

- paraphrasing
- deletion
- splitting

evaluation criteria

- grammaticality
- meaning preservation
- simplicity

not easy to measure

System Comparability

We need more controlled evaluation:

We need more controlled evaluation:

grammaticality
meaning preservation
simplicity

Not easy to measure

paraphrasing
deletion
splitting

System Comparability

We need more controlled evaluation:
- evaluate sub-tasks separately

not easy to measure
System Comparability

We need more controlled evaluation:
- evaluate sub-tasks separately
- target specific audience (e.g. 10-12 year old)

Opinion #2

Simple Wikipedia is **not** that simple
Photolithography

Microphotolithography is the use of photolithography to transfer geometric shapes on a photomask to the surface of a semiconductor wafer for making integrated circuits.
“Specific questions that need addressing are:

… we need to better understand the quality of Simple English Wikipedia, a resource that has been used to train many SMT based simplification systems…”

—— Advaith Siddharthan (2014 Survey)
“Specific questions that need addressing are:

… we need to better understand the quality of Simple English Wikipedia, a resource that has been used to train many SMT based simplification systems…”

—— Advait Siddharthan (2014 Survey)

**WHAT’S NEW?**

We quantitively and systematically answer this quest.

Quality of Parallel Wikipedia Corpus

Alignment error: 17%
Real simplification: 50%
Not simpler: 33%
Inaccuracy in Parallel Wikipedia Corpus*

alignment error (two sentences have different meaning)

17%
Inaccuracy in Parallel Wikipedia Corpus*

alignment error (two sentences have different meaning)

17%

Best automatic sentence alignment gets about 0.7 F1 score (Hwang et al. 2015)

Inadequacy in Parallel Wikipedia Corpus

**alignment error**
(two sentences have different meaning)

Best automatic sentence alignment gets about **0.7 F1** score (Hwang et al. 2015)

Inadequacy in Parallel Wikipedia Corpus*

**alignment error**

(two sentences have different meaning)

Best automatic sentence alignment gets about **0.7 F1** score (Hwang et al. 2015)

**not simpler**

Sentences can have similar meaning but not simplification

---

Inadequacy in Parallel Wikipedia Corpus*

real simplification (aligned and simpler)

50%
Inadequacy in Parallel Wikipedia Corpus

real simplification (aligned and simpler)

50%

Inadequacy in Parallel Wikipedia Corpus*

real simplification (aligned and simpler)

Some sentences are simpler by only one word while the rest of sentence is still complex

Issues with Parallel Wikipedia Corpus

Issues with Parallel Wikipedia Corpus

• suboptimal for estimating “translation” probabilities
Issues with Parallel Wikipedia Corpus

- suboptimal for estimating "translation" probabilities
- suboptimal for developing automatic metrics
Issues with Parallel Wikipedia Corpus

- suboptimal for estimating “translation” probabilities
- suboptimal for developing automatic metrics
- suboptimal for tuning MT system
Issues with Parallel Wikipedia Corpus

- suboptimal for estimating “translation” probabilities
- suboptimal for developing automatic metrics
- suboptimal for tuning MT system
- unsuitable for document-level simplification
Opinion #3

New data *can* help
Newsela Dataset

every article at 5 levels of simplification
written by trained editors, comes with comprehension quizzes

Wikipedia*

manual inspection of aligned sentence pairs

alignment error: 17%
real simplification: 50%
not simpler: 33%

Newsela

alignment error: 2%
real simplification: 92%
not simpler: 6%

Good simplification needs more paraphrasing.

Wikipedia* Newsela

Good simplification could be much shorter.

sentence length (#words)

see syntax analysis in the paper

Good simplification uses a much smaller vocabulary.

Wikipedia* (total 2.6 million tokens)

- Normal vocabulary size: 71,340
- Simple vocabulary size: 6,669
- 18% reduction

Newseela (total 1.3 million tokens)

- Normal vocabulary size: 19,849
- Simple vocabulary size: 19,197
- 48% reduction

Good simplification reduces certain function word usage.

- commune
- ,
- as
- and
- northern
- northwestern
- film
- ;
- southwestern
- footballer

most significantly reduced words
(weighted log-odds-ratio analysis w/ informative Dirichlet prior)

Postal officials recently tried to … … , which could … ….

Postal officials recently tried to … … . That could … ….

most significantly reduced words

see syntax analysis in the paper

Postal officials recently tried to … … , **which** could … ….

**That** could … ….

most significantly reduced words

see syntax analysis in the paper

Postal officials recently tried to … … , which could … … . That could … … .

most significantly reduced words
see syntax analysis in the paper

Wikipedia is not suitable for full-document simplification.

document compression ratio (simple/normal)
see discourse analysis in the paper

Opinion #1  Current evaluation doesn’t tell us what’s going on.

Opinion #2  Simple Wikipedia is not that simple.

Opinion #3  New data can help.
My Suggestions
My Suggestions

• to reviewers:
My Suggestions

- to reviewers:
  - be open-minded to papers that may not follow previous evaluation setup, may not outperform the “state-of-the-art” on Wikipedia
My Suggestions

- **to reviewers:**
  - be open-minded to papers that may not follow previous evaluation setup, may not outperform the “state-of-the-art” on Wikipedia
  - be sympathetic towards papers specially on data construction*, data analysis* and automatic evaluation metrics

* (Pellow & Maxine, 2014 HCOMP; Marcelo & Specia, 2014 PITR)

My Suggestions

• to reviewers:
  - be open-minded to papers that may not follow previous evaluation setup, may not outperform the “state-of-the-art” on Wikipedia
  - be sympathetic towards papers specially on data construction*, data analysis* and automatic evaluation metrics
  - read our paper

* (Pellow & Maxine, 2014 HCOMP; Marcelo & Specia, 2014 PITR)

My Suggestions
My Suggestions

- to researchers:

My Suggestions

• to researchers:

- consider working on text simplification ("pre-BLEU age")
My Suggestions

- **to researchers:**
  - consider working on text simplification ("pre-BLEU age")
  - improve evaluation
My Suggestions

• to researchers:
  - consider working on text simplification ("pre-BLEU age")
  - improve evaluation
  - make your system replicable

My Suggestions

• to researchers:
  - consider working on text simplification ("pre-BLEU age")
  - improve evaluation
  - make your system replicable
  - read our paper

Thank you

Questions?
Opinions?

Sponsor: NSF
Newsela data are available at https://newsela.com/data/

Back Up
Reasons of Quality Issues in Parallel Wikipedia Corpus

- The Simple Wikipedia was created by volunteers with no specific objective;
- Articles in Simple Wikipedia do not necessarily map Normal Wikipedia;
- As an encyclopedia, Wikipedia contains extremely difficulty words and sentences.
| Simple-1 | Fourth-graders in most states are better readers than they were a decade ago. But only a third of them actually are able to read well, according to a new report. |
| Simple-2 | Fourth-graders in most states are better readers than they were a decade ago. But only a third of them actually are able to read well, according to a new report. |
| Simple-3 | Most fourth-graders are better readers than they were 10 years ago. But few of them can actually read well. |
| Simple-4 | Fourth-graders are better readers than 10 years ago. But few of them read well. |

Original: Slightly more fourth-graders nationwide are reading proficiently compared with a decade ago, but only a third of them are now reading well, according to a new report.
Newsela Dataset

1,130 news articles

Time: 2013 January ~ 2015 March

Source: Chicago Tribune, Seattle Times, LA Times, The Baltimore Sun

Original: 56k sentences — Simple: 64k sentences