# Optimizing Statistical Machine Translation for Text Simplification

Wei Xu[1,3], Courtney Napoles[2], Ellie Pavlick[1], Quanze Chen[1], and Chris Callison-Burch[1]
[1] *University of Pennsylvania,* [2] *Johns Hopkins University,* [3] *Ohio State University*

## Introduction

- A state-of-the-art simplification system
- An automatic evaluation metric: ***SARI***
- Two multi-reference simplification datasets

## 1. Paraphrase Grammars

We use the Paraphrase Database (PPDB) as our paraphrase rules [Ganitkevitch et al. 2013]:

- >1000X larger than the parallel Wikipedia corpus
- More diverse paraphrases
- Many types of transformations

*Lexical:*

**solely** ➔ **only**

*Phrasal:*

**generally acknowledged** ➔ **widely accepted**

*Syntactic:*

**the manner in which NN** ➔ **the way NN**

## 2. Simplification Features

We augment the paraphrase rules with simplification-specific features, specifically:

- character and word length
- number of syllables
- Gigaword and Simple Wikipedia LM scores
- # common words / total words

## Want to learn more? Read our TACL papers!

Xu et al. 2016. Optimizing Statistical Machine Translation for Simplification. TACL.

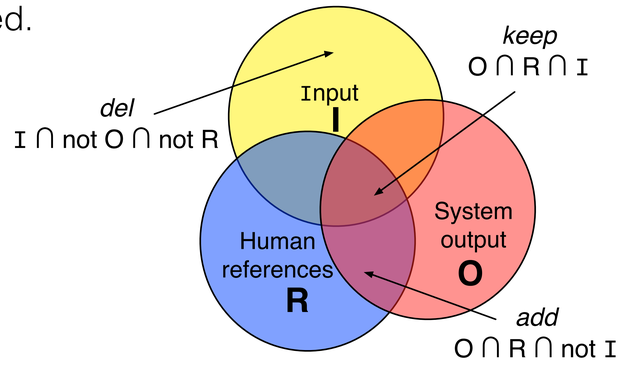
Xu et al. 2015. Problems in Current Text Simplification Research: New Data Can Help. TACL.

## 3. Metrics

We designed two tunable metrics for pairwise ranking optimization (PRO) [Hopkins and May 2011] in the Joshua MT Pipeline [Post et al. 2013].

### *SARI*

Compare **S**ystem output **a**gainst **R**eference sentences and against the **I**nput sentence. Measure goodness of n-grams that system kept, added, or deleted.




$$\text{SARI} = d_1 F_{add} + d_2 F_{keep} + d_3 P_{del}$$
$$d_1 = d_2 = d_3 = 1/3$$

### *FKBLEU*

Combination of the FK grade level and iBLEU, which rewards O ∩ R and penalizes O ∩ I

$$\text{FKBLEU} = \text{iBLEU}(I, R, O)^{1/2} \times \text{sigmoid}(\text{FK}(O) - \text{FK}(I))^{1/2}$$

## 4. Multiple Reference Simplifications

- Wikipedia and Simple English Wikipedia
- 2000 sentences for tuning, 350 for evaluation
- We collected additional, manual simplifications (8 references) on Mechanical Turk.

## Data and Code

**https://cocoxu.github.io**

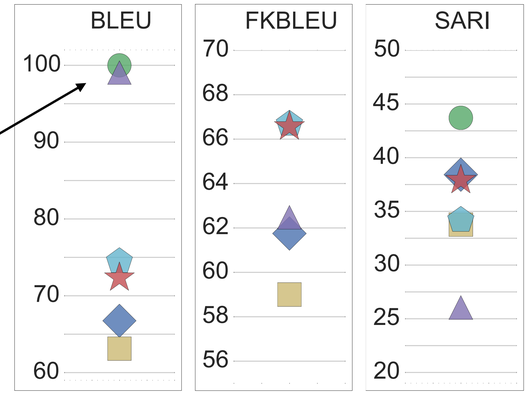
Newsela Simplification Corpus (1000+ news articles)

## 5. Results

We evaluated our system, SBMT-X (X is the tuning metric), against these simplifications:

- Simple Wikipedia
- Mechanical Turk
- PBMT-R [Wubben et al. 2012]
- SBMT-BLEU
- SBMT-FKBLEU
- SBMT-SARI

### *Automatic metrics*




**Problem of BLEU** outputs with no change get high BLEU and high Grammar/Meaning

Range of all metrics: 0-100.

### *Correlation with human ratings*

| *Spearman's rho* | *Grammar* | *Meaning* | *Simplicity* |
|---|---|---|---|
| *BLEU* | 0.59 | 0.70 | 0.15 |
| *FKBLEU* | 0.35 | 0.41 | 0.24 |
| ***SARI*** | **0.34** | **0.40** | **0.34** |

### *Human evaluation*

*Grammar:* How grammatical? (0-4)
*Meaning:* How much original meaning? (0-4)
*Simplicity+:* # of good paraphrases (0-Inf.)