Human Evaluation for Text Simplification: The Simplicity-Adequacy Tradeoff

Max Schwarzer
Computer Science Department
Pomona College
max.schwarzer@pomona.edu

David Kauchak
Computer Science Department
Pomona College
david.kauchak@pomona.edu

Abstract

In this paper we examine human evaluation for text simplification. We find a strong inverse correlation between simplicity and adequacy, hinting that caution should be used when comparing systems across these metrics. Additionally, we examine the impact of test set size and the number of human annotators, finding that test set size is critical while using multiple human annotators has only limited benefit.

1 Introduction

There are currently two main types of methods used to assess text simplification systems. The first are automated metrics such as BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016) which compare system output to a human generated simplification and measure how similar the two are. The second are human evaluation metrics where the output of the system is judged by human annotators, usually along three dimensions: *simplicity*, measuring how simple the text looks; *fluency*, measuring how much the output looks like fluent, grammatically correct text; and *adequacy*, measuring how well the content in the original, unsimplified sentence is preserved in the simplified text. Test sentences are scored independently by multiple annotators often using crowdsourcing platforms such as Amazon's Mechanical Turk (Callison-Burch and Dredze, 2010).

Automated metrics are important for development, tuning and quick analysis, but current metrics do not perfectly capture the varied dimensions required for evaluating text simplification, e.g. BLEU tends to correlate with adequacy, while SARI tends to correlate with simplicity and fluency (Štajner et al., 2014; Xu et al., 2016). Because of this, human metrics still play an important role in understanding and comparing the performance of text simplification systems. In this paper, we provide additional quantitative analysis to understand how current human evaluation metrics interact, and examine how test set size and the number of annotators impacts evaluation discrimination.

2 Experimental Setup

We evaluated three different text simplifications systems. We included two phrase-based approaches, Moses-Del (Coster and Kauchak, 2011a) and PBMT-R (Wubben et al., 2012), and one syntax-based approach, SimpleTT (Feblowitz and Kauchak, 2013). We used the sentence-aligned Wikipedia corpus from Coster and Kauchak (Coster and Kauchak, 2011b) and randomly selected 100 sentence pairs from the test portion of this dataset. We also included the original, unsimplified sentence (English Wikipedia) and the human simplified variant (Simple English Wikipedia) in our tests, resulting in a dataset of 500 sentences (five variants of 100 sentences) of varying quality and characteristics.

We measured the quality of these simplifications based on three human evaluation metrics, *simplicity* (-2 to 2 scale), *fluency* (1 to 5 scale) and *adequacy* (1 to 5 scale), all of which have been frequently used previously and are the standard metrics for human evaluation (Xu et al., 2016). For each of these three metrics, for each of the 500 sentences, we collected annotations from 10 annotators using Amazon Mechanical Turk (MTurk). We averaged results across annotators for each sentence, resulting in 500 data points for each of the three metrics.

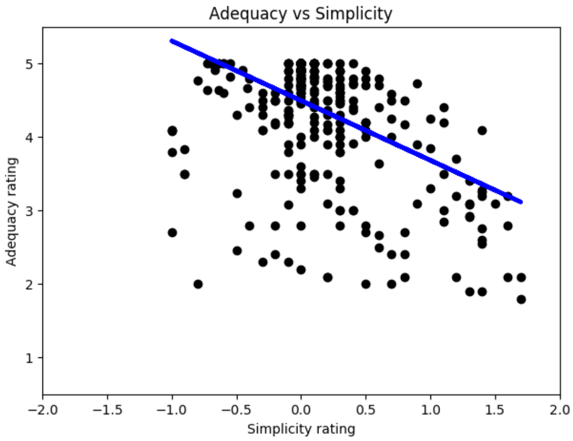

Figure 1: Simplicity vs adequacy for all the sentences in the test set

3 Results

3.1 Correlations between metrics

We noticed two correlations between metrics; simplicity is negatively correlated with adequacy ($R^2 = 0.21$, $p \ll 0.0001$), and adequacy is positively correlated with fluency ($R^2 = 0.07$, $p \ll 0.0001$). There was no significant correlation between simplicity and fluency. Figure 1 shows a plot of the relationship between simplicity and adequacy with line of best fit drawn.

3.2 Test set size

To investigate how large a test set is required, we randomly sampled 1,000 samples with replacement from our original dataset to create new datasets of increasing size ranging from 10 sentences to 100,. We then measured the degree to which the results of the (smaller) resamples were correct by comparing the ranked pairs of the five approaches to the full test set. We measured if two systems were different using a paired t-test ($p < 0.01$) and then measured agreement using Kendall's Tau. We present the results here for adequacy, though similar trends were seen for the other metrics.

Figure 2 shows the plot of Kendall's Tau vs. the test set size. For small test sets, we can frequently either make the wrong conclusion or are unable to draw the correct conclusion. With test set sizes around 50-60 sentences the performance levels off, though it still continues to increase slightly.

A related question is how many annotators should be used. To test this, we kept the sentence sized fixed, but randomly sampled with replacement a new dataset with increasing number of annotators (from 1 to 10). Even with just a single annotator (though not the *same* annotator for all 100 samples) we achieve strong agreement with a $\tau = 0.92$ and by increasing to just two annotators we reach a peak of $\tau = 0.95$. After that, adding additional annotators does not significantly improve the discriminating power.

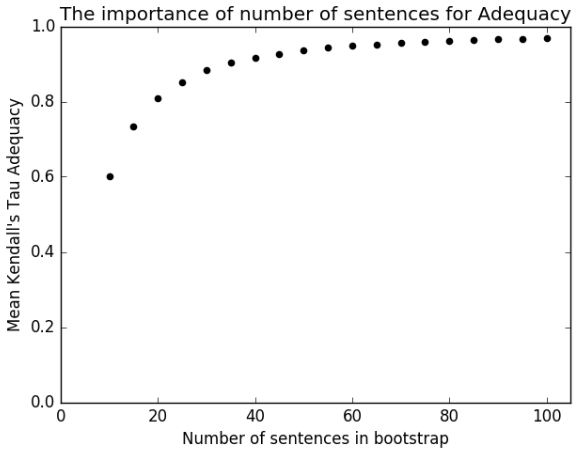

Figure 2: Kendall's Tau on adequacy for 100 sentences with increasing number of annotators per sentence.

Finally, we examined the situation where only a fixed number of annotations are available (e.g., due to financial constraints). We fixed the total number of annotations at 100, and tested every possible test size with 1 to 10 annotators per sentence (rounding down when necessary). We found that ranking agreement monotonically decreased as more annotators were used per sentence, indicating that test set size should be prioritized.

4 Discussion

Our finding that adequacy and simplicity are negatively correlated suggests a common, underlying fact: removing material from a sentence will make it simpler, while reducing its adequacy. Critically, this correlation suggests that caution should be used when comparing systems across all three metrics: *improvement in one metric and not the other may be due to this inverse relationship rather than actual system performance.*

For test set size, we suggest that using multiple annotators per sentence, particularly beyond two, is not beneficial for human evaluation. Instead, more sentences should be evaluated to improve the robustness of the results.

References

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.*

William Coster and David Kauchak. 2011a. Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*.

William Coster and David Kauchak. 2011b. Simple English Wikipedia: A new text simplification task. In *Proceedings of ACL.*

Daniel Feblowitz and David Kauchak. 2013. Sentence simplification as tree transduction. In *Proceedings of PITR (ACL Workshop).*

K. Papineni, S. Roukos, T. Ward, W.J., and Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL.*

Sanja Štajner, Ruslan Mitkov, and Horacio Saggion. 2014. One step closer to automatic evaluation of text simplification systems. In *Proceedings PITR@EACL.*

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of ACL.*

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*.