A Closer Look at Recent Results of Verb Selection for Data-to-Text NLG (Sumplementary Materials)

Guanyi Chen
Utrecht University
g.chen@uu.nl

Jin-Ge Yao
Microsoft Research Asia
$jinge \cdot yao@microsoft \cdot com$

A Instructions for Human Judgements

In each question, an AMT worker will see the following instruction:

Below are 11/14 sentences describing an upward/downward percentage change using different verbs. For each sentence, please select one of the three degrees of appropriateness:

- 3 (Appropriate): The verb is among the most suitable ones to describe the percentage
- 2 (Okay): The verb could be used to describe the percentage, although it might not be one of the most appropriate choices
- 1 (Not Appropriate): The verb is not naturally used to describe the percentage

Figure 1 show an example screenshot.

B Cross-entropy

Besides normal metrics for classification, we also calculated the cross-entropy for different output distributions compared with human annotated data, with results displayed in Table 1.

C Evaluation Details and Example Outputs

The models were evaluated in a way as a multi-label classification task for which the macro-averaged precision, recall, and F-score are used. The accumulated thresholds were tune once the model achieved best precision score on the development set. Table 2 first lists a number of examples of human choices given different percentage changes. Based on those same changes, we asked different models to select verbs and listed them in the rest of Table 2 together with the turned thresholds.

Model	upward verds		downward verbs	
	$CE_a \downarrow$	$CE_n \uparrow$	$CE_a \downarrow$	$CE_n \uparrow$
Thomson Reuters	5.54	6.23	5.71	5.75
Neural Network	2.88	4.48	3.31	3.98
Frequency	3.01	4.31	3.16	4.13
Decision Tree	2.88	4.28	3.16	4.13
Bayesian ($\lambda = 1$, KDE)	2.90	4.73	3.15	4.12
Bayesian ($\lambda = 1$, Beta)	3.00	5.13	3.23	4.17
Bayesian ($\lambda = 0.05$, KDE)	2.48	3.05	2.69	2.87
Bayesian ($\lambda = 0.05$, Beta)	2.68	3.87	2.82	3.02

Table 1: The average cross entropy for both human judged appropriate verbs and inappropriate verbs; ↑ means the value of the metric is the higher the better, while ↓ means the lower the better.

Model	Selected Verbs
human	1.86%: gain, grow, increase, rise 33.87%: grow, soar, advance, climb, gain, increase, jump, rise 93.19%: rise, increase, raise, soar, surge

Thomson Reuters
$\gamma = 0.501$
	advance(0.00), boost(0.00), climb(0.00), gain(0.00), grow(0.00), increase(0.00), jump(0.00), raise(0.00), rise(0.00), soar(0.00), surge(0.00)
	jump(0.33), soar(0.33)
	advance(0.00), boost(0.00), climb(0.00), gain(0.00), grow(0.00), increase(0.00), jump(0.00), raise(0.00), rise(0.00), soar(0.00), surge(0.00)

Neural Network
$\gamma = 0.982$
	rise(0.61), increase(0.16), grow(0.09), climb(0.05), gain(0.02), jump(0.02), raise(0.01), surge(0.01))
	rise(0.54), increase(0.15), jump(0.08), grow(0.06), climb(0.06), surge(0.04), soar(0.03), gain(0.02), raise(0.01)
	rise(0.47), jump(0.17), increase(0.11), soar(0.10), surge(0.07), climb(0.03), grow(0.03)

Frequency $\gamma = 0.993$	rise(0.59), increase(0.16), grow(0.08), climb(0.05), jump(0.04), surge(0.02), gain(0.02), soar(0.01), raise(0.01), advance(0.01), boost(0.01) rise(0.59), increase(0.16), grow(0.08), climb(0.05), jump(0.04), surge(0.02), gain(0.02), soar(0.01), raise(0.01), advance(0.01), boost(0.01) rise(0.59), increase(0.16), grow(0.08), climb(0.05), jump(0.04), surge(0.02), gain(0.02), soar(0.01), raise(0.01), advance(0.01), boost(0.01)

Decision Tree
$\gamma = 0.983$
	rise(0.61), increase(0.16), grow(0.08), climb(0.05), jump(0.02), gain(0.02), surge(0.02), raise(0.01), advance(0.01)
	rise(0.53), increase(0.15), jump(0.08), grow(0.07), climb(0.05), surge(0.04), soar(0.03), gain(0.02), raise(0.01), boost(0.01)
	rise(0.53), increase(0.15), jump(0.08), grow(0.07), climb(0.05), surge(0.04), soar(0.03), gain(0.02), raise(0.01), boost(0.01)

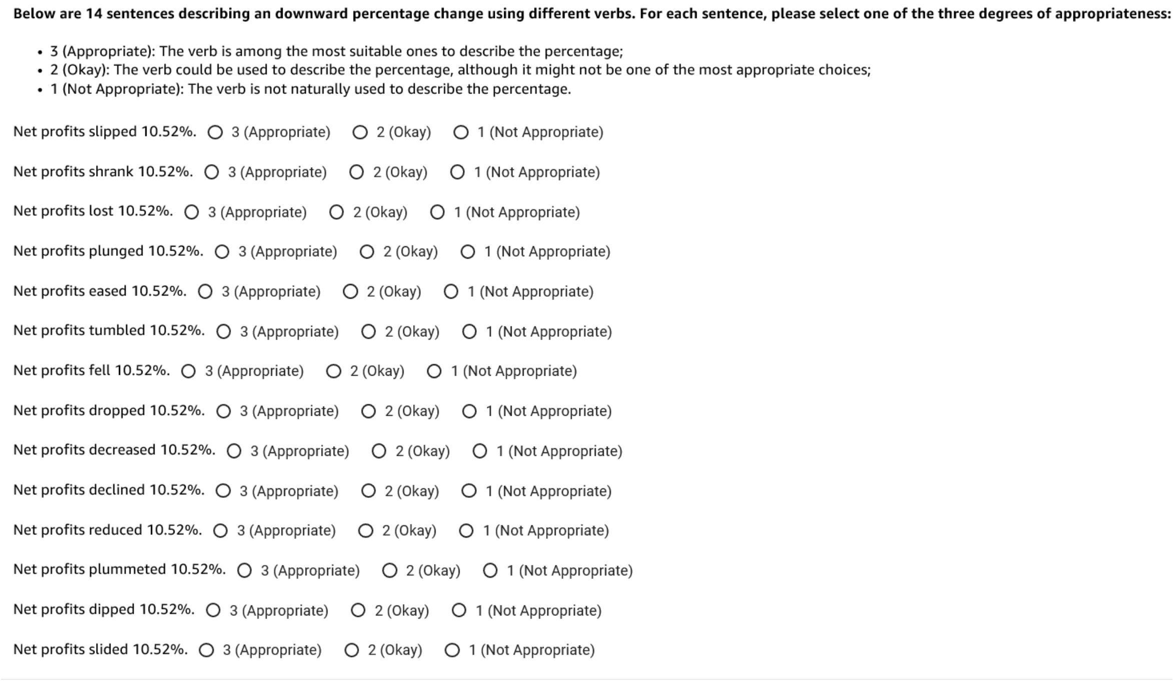
Bayesian
($\lambda = 1$, KDE)
$\gamma = 0.977$
	rise(0.64), increase(0.15), grow(0.09), climb(0.04), gain(0.02), jump(0.02), surge(0.01), raise(0.01)
	rise(0.55), increase(0.15), grow(0.07), jump(0.07), climb(0.05), surge(0.03), soar(0.02), gain(0.02), raise(0.02)
	rise(0.44), jump(0.15), soar(0.12), increase(0.09), surge(0.09), climb(0.05), grow(0.03)

Bayesian
($\lambda = 1$, Beta)
$\gamma = 0.911$
	rise(0.62), increase(0.16), grow(0.08), climb(0.05)
	rise(0.57), increase(0.16), grow(0.07), jump(0.06), climb(0.05), surge(0.03)
	soar(0.35), jump(0.27), surge(0.20), rise(0.13)

Bayesian
($\lambda = 0.05$, KDE)
$\gamma = 0.496$
	rise(0.16), grow(0.13), gain(0.12), increase(0.12)
	jump(0.12), soar(0.12), surge(0.11), rise(0.09), raise(0.09)
	soar(0.42), surge(0.20)

Bayesian
($\lambda = 0.05$, Beta)
$\gamma = 0.623$
	rise(0.14), gain(0.13), grow(0.12), increase(0.11), climb(0.11), advance(0.10)
	boost(0.11), jump(0.11), rise(0.11), soar(0.10), surge(0.10), increase(0.09), raise(0.09)
	soar(0.60), surge(0.22)

Table 2: The first row shows examples of human annotated appropriate, okey, and not appropriate trend verbs given certain percentage changes. Recall that the verbs appear in this table were filtered by majority voting, i.e., annotated by more than three (out of five) subjects. The rest rows list verbs selected by each model, together with corresponding probabilities, where γ is its tuned threshold, the selected verbs are generated with respected to the same inputs as the human choice examples (presenting in the same order).



Below are 14 sentences describing an downward percentage change using different verbs. For each sentence, please select one of the three degrees of appropriateness:

- 3 (Appropriate): The verb is among the most suitable ones to describe the percentage;
- 2 (Okay): The verb could be used to describe the percentage, although it might not be one of the most appropriate choices;
- 1 (Not Appropriate): The verb is not naturally used to describe the percentage.

Net profits slipped 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits shrank 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits lost 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits plunged 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits eased 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits tumbled 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits fell 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits dropped 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits decreased 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits declined 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits reduced 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits plummeted 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits dipped 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Net profits slided 10.52%. ○ 3 (Appropriate) ○ 2 (Okay) ○ 1 (Not Appropriate)

Figure 1: The screen shot of a sample question for our experiment