Svetlana Kiritchenko and Saif M. Mohammad
National Research Council Canada
Svetlana.Kiritchenko@nrc-cnrc.gc.ca, Saif.Mohammad@nrc-cnrc.gc.ca

NRC·CNRC Information and Communications Technologies

Best–Worst Scaling More Reliable than Rating Scales:

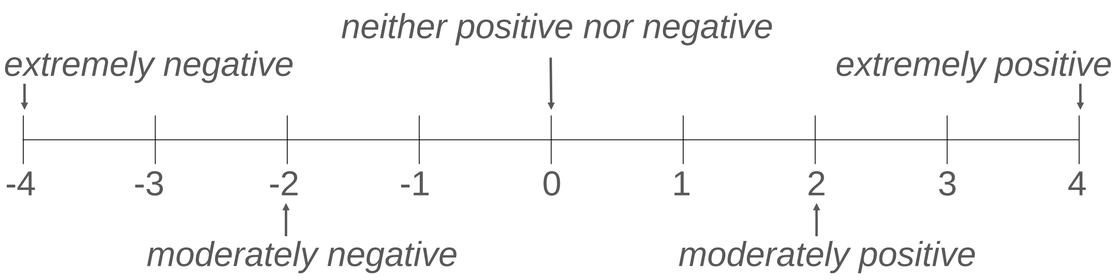
A Case Study on Sentiment Intensity Annotation

In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-2017), August, 2017, Vancouver, Canada.

1. Fine-Grained Dimensional Annotation

1.1 Rating Scales (Traditional Method)

Annotation question: Rate an item on a scale (e.g., *strongly disagree* to *strongly agree*, *wickedly yucky* to *wickedly yummy*)



Obtaining real-valued scores: annotations for a term from multiple respondents are averaged.

Problems with Rating Scales (RS):
- inconsistencies in annotations by different annotators
- inconsistencies in annotations by the same annotator
- scale region bias
- fixed granularity

1.2 Best–Worst Scaling (Louviere and Woodworth, 1990)

Annotation questions: Given a 4-tuple (4 items),
- which item is the **Best** (e.g., the *most positive*)?
- which item is the **Worst** (e.g., the *most negative*)?

most negative	4-tuple	most positive
◯	*violence*	◯
◯	*increasingly attractive*	◯
◯	*permission*	◯
◯	*will not be interested*	◯

Obtaining real-valued scores (Orme, 2009):

$$score(t) = \%best(t) - \%worst(t)$$

Advantages of Best–Worst Scaling (BWS):
- addresses RS problems through item comparisons
- good results with annotations for ~2N 4-tuples
  - multiple sets of 2N tuples generated randomly
  - set that maximizes tuple diversity is chosen

2. Quantitative Comparison of Rating Scale and Best–Worst Scaling

Hypothesis: BWS produces more reliable ranking than rating scales for the same total number of annotations.

Experimental set-up:
- We annotate 3,207 (N) English terms (words and phrases) by crowdsourcing
- RS: Each of the N terms is labeled by 20 respondents
- BWS: Each of the 2N 4-tuples is labeled by 10

Quantitative comparison:

Q1. How different are the annotations?

Q2. How reproducible are the term scores and rankings?

Q1. Differences in outcomes of RS and BWS

Differences in final outcomes of BWS and RS, for different total numbers of annotations (N=3,207 is the number of terms)

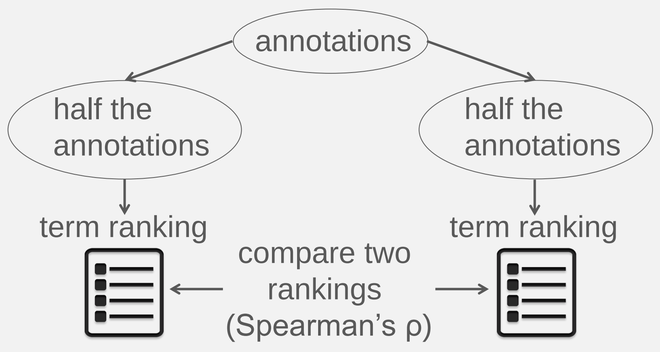
# annotations	avg. Δ score (0..1)	avg. Δ rank	ρ	r
3N	0.11	397	0.85	0.85
5N	0.10	363	0.87	0.88
20N	0.08	264	0.93	0.93

Conclusions: the ranks/scores diverge considerably, especially for commonly used annotation scenarios with only 3N or 5N total annotations.

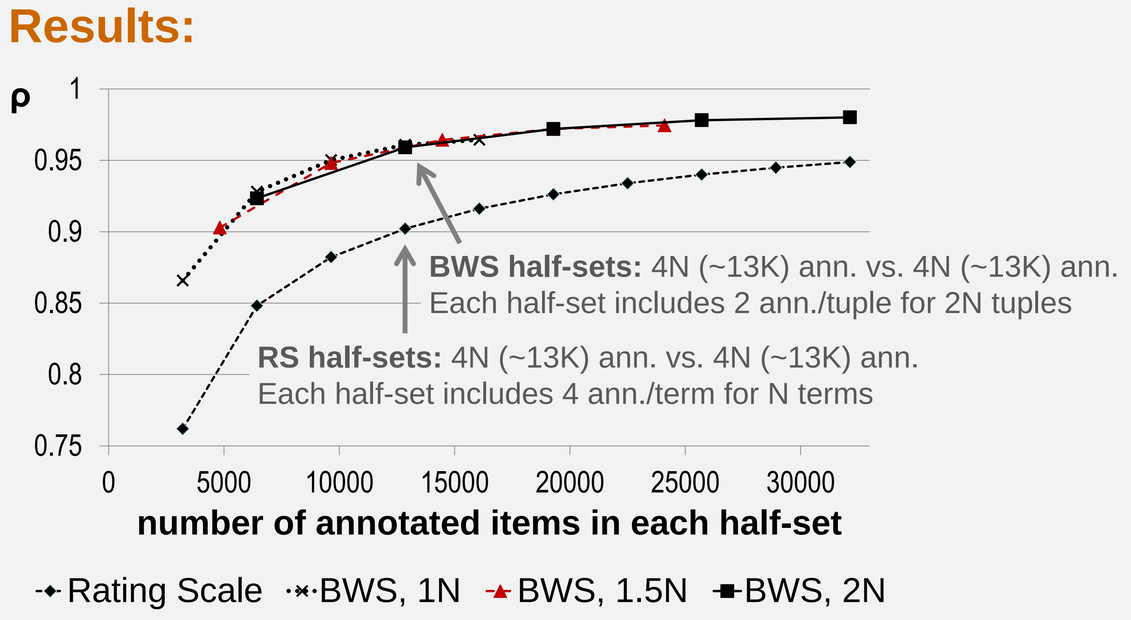
Q2. Reproducibility

If repeated annotations from multiple respondents result in similar sentiment scores, then one can be confident that the scores capture the true sentiment intensities.

Split-half reliability:



Results:



Conclusions:
- BWS surpasses RS on the ability to reliably rank items by sentiment, especially for phrasal items.
- The reliability obtained by RS with 10 annotations/term is matched by BWS with only 3N total annotations.

All data and scripts used in this project are available at:
http://www.saifmohammad.com/WebPages/bwsVrs.html

Code for Best–Worst Scaling and all lexicons are available at:
http://www.saifmohammad.com/WebPages/BestWorst.html