Sequence Effects in Crowdsourced Annotations

Nitika Mathur **Timothy Baldwin** **Trevor Cohn**
School of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia
nmathur@student.unimelb.edu.au
{tbaldwin,tcohn}@unimelb.edu.au

Abstract

Manual data annotation is a vital component of NLP research. When designing annotation tasks, properties of the annotation interface can lead to unintentional artefacts in the resulting dataset, biasing the evaluation. In this paper, we explore sequence effects where annotations of an item are affected by the preceding items. Having assigned one label to an instance, the annotator may be less (or more) likely to assign the same label to the next. During rating tasks, seeing a low quality item may affect the score given to the next item either positively or negatively. We see clear evidence of both types of effects using auto-correlation studies over three different crowdsourced datasets. We then recommend a simple way to minimise sequence effects.

1 Introduction

NLP research relies heavily on annotated datasets for training and evaluation. The design of the annotation task can influence the decisions made by annotators in subtle ways: besides the actual features of the instance being annotated, annotators are also influenced by factors such as the user interface, wording of the question, and familiarity with the task or domain.

When collecting NLP annotations, care is usually taken to ensure that the annotations are of high quality, through careful design of label sets, annotation guidelines and training of annotators (Hovy et al., 2006), methods for aggregating annotations (Passonneau and Carpenter, 2014), and intuitive user interfaces (Stenetorp et al., 2012).

Crowdsourcing has emerged as a cheaper, faster alternative to expert NLP annotations (Snow et al., 2008; Callison-Burch and Dredze, 2010; Graham et al., 2017), although it entails additional effort to filter out unskilled or opportunistic workers, e.g. through the collection of redundant repeated judgements for each instance, or including some trap questions with known answers (Callison-Burch and Dredze, 2010; Hoßfeld et al., 2014). In most annotation exercises, the order of presentation of instances is randomised to remove bias due to similarities in topic, style and vocabulary (Koehn and Monz, 2006; Bojar et al., 2016).

When crowdsourcing judgements, the normal practise (as used in the datasets we analyse) is for the item ordering to be randomised in creating a "HIT" (i.e. a single collection of items presented to a crowdworker for judgement), and then to have each HIT annotated by multiple workers, for quality control purposes. The order of items is generally fixed across all annotators of an individual HIT (Snow et al., 2008; Graham et al., 2017).

In this paper, we show that worker scores are affected by sequence bias, whereby the order of presentation can affect individuals' assessment of an item. Since all workers see the instances in the same order, this affects any other inferences made from the data, including aggregated assessment or inferences about individual annotators (such as their overall quality or individual thresholds).

Possible explanations for sequence effects include:

Gambler's fallacy: Once annotators have developed an idea of the distribution of scores/labels, they can come to expect even small sequences to follow the distribution. In particular, in binary annotation tasks, if they expect that True (1) and False (0) items are equally likely, then they believe the sequence 00000 (100% False and 0% True) is less likely than the sequence 01010 (50% False and 50% True). So if they assign 0 to an item,

they may approach the next item with a prior belief that it is more likely to be a 1 than a 0. Chen et al. (2016) showed evidence for the gambler's fallacy in decisions of loan officers, asylum judges, and baseball umpires.

Sequential contrast effects: A high quality item may raise the bar for the next item. On the other hand, a bad item may make the next item seem better in comparison (Kenrick and Gutierres, 1980; Hartzmark and Shue, to appear)

Assimilation and anchoring: The annotator uses their score of the previous item as an anchor, and adjusts the score of the current item from this anchor, based on perceived similarities and differences with the previous item. If they focus on similarities between the previous and current instance, the annotations show an assimilation effect (Geiselman et al., 1984; Damisch et al., 2006). Anchoring effects may decrease as people gain experience and expertise in the task (Wilson et al., 1996).

2 Methodology

We test whether the annotation of an instance is correlated with the annotation on previous instances, conditioned on control variables such as the gold standard (i.e. expert annotations[1]), based on the following linear model:

$$\mathrm{Y}_{i,t} = \beta_0 + \beta_1 \mathrm{Y}_{i,t-1} + \beta_2 \mathrm{Gold} + \eta \qquad (1)$$

where $\mathrm{Y}_{i,t}$ is the annotation given by an annotator i to an instance t, and η is white Gaussian noise with zero mean. We use linear regression for continuous data and logistic regression for binary data.[2] If there is no dependence between consecutive instances, and annotators assign labels/scores based only on the aspects of the current instance, then the data can be explained from the gold score (learning a positive β_2 value) and bias term (β_0), with β_1 set to zero. When we use the ground truth as a control, if β_1 is non-zero, it is evidence of mistakes being made by annotators due to sequential bias. A positive value of β_1 can be explained by priming or anchoring, and a negative value with sequential contrast effects or the gambler's fallacy. Accordingly, we test the statistical significance of the $\beta_1 \neq 0$ to determine whether sequencing effects are present in crowdsourced text corpora.

[1] For the Machine Translation dataset described in Section 3.3, we use the mean of at least fifteen crowd workers as a proxy for expert annotations.

[2] η is not included in the case of logistic regression

Task	All	Good	Moderate
RTE	-0.102	-0.169^{*}	-0.192^{**}
TEMPORAL	0.198	-0.567^{***}	-0.511^{***}

Table 1: Autocorrelation coefficient β_1 for RTE and TEMPORAL data. Stars denote statistical significance: * = 0.05, ** = 0.01, and *** = 0.001.

3 Experiments

We analyse several influential datasets that have been constructed through crowdsourcing, including both binary and continuous annotation tasks: recognising textual entailment, event ordering, affective text analysis, and machine translation evaluation.

3.1 Recognising Textual Entailment (RTE) and Event Temporal Ordering

First, we examine the recognising textual entailment ("RTE") and event temporal ordering ("TEMPORAL") datasets from Snow et al. (2008). In the RTE task, annotators are presented with two sentences, and are asked to judge whether the second text can be inferred from the first. With the TEMPORAL dataset, they are shown two sentences describing events, and asked to indicate which of the two events occurred first. Both datasets include both expert annotations and crowdsourced annotations constructed using Amazon Mechanical Turk ("MTurk"). On MTurk, each RTE HIT contains 20 instances, and each TEMPORAL HIT contains 10 instances, which the workers see in sequential order. For both tasks, each HIT was annotated by 10 workers.

Results We use logistic regression on worker labels against labels on the previous instance in the current HIT, with the expert judgements as a control variable. We also add an additional control, namely the percentage of True labels assigned by the worker overall, which accounts for the overall annotator bias. To calculate this, we use scores by the worker excluding the current score, to avoid giving the model any information about the current instance.

As shown in Table 1, over all workers ("All"), we find a small negative autocorrelation for both the RTE and TEMPORAL tasks. One possibility

is that this is biased by opportunistic workers who assign the same label to all instances in the HIT, for which we would not expect any sequential bias effects. When we exclude these workers ("Moderate"), the autocorrelation increases, and is highly statistically significant. We also show results for workers with at least 60% accuracy when compared to expert annotations ("Good"), and observe a similar effect.

3.2 Affective text analysis

In the affective text analysis task ("AFFECTIVE"), annotators are asked to rate news headlines for anger, disgust, fear, joy, sadness, and surprise on a continuous scale of 0–100. Besides these emotions, they are asked to rate sentences for (emotive) valence, i.e., how strongly negative or positive they are (-100 to $+100$). In this dataset, there are 100 headlines divided into 10 HITs, with 10 workers annotating each HIT (Snow et al., 2008). We test for autocorrelation of scores of each aspect individually, controlling for the expert scores and worker correlation with the expert scores. We also look separately at datasets of good and bad workers, based on whether the correlation with the expert annotations is greater than 0.5.

Results For individual emotions, we do not observe any significant autocorrelation ($p \geq 0.05$). As there are only 1000 annotations per emotion, we also look at results when combining data for all aspects. Though we find a statistically significant negative autocorrelation for scores of the full dataset, this disappears when we filter out bad workers (Table 2). Given the difficulty of this very subjective task, it is likely that many of workers considered 'bad' might have simply found this task too difficult or arbitrary, and thus become more prone to sequence effects.

3.3 Machine Translation Adequacy

When evaluating machine translation ("MT"), we tend to focus on adequacy: the extent to which the meaning of the reference translation is captured in the MT output. In the method of Graham et al. (2015) — the current best-practise, as adopted by WMT (Bojar et al., 2016) — annotators are asked to judge the adequacy of translations using a 100-point sliding scale which is initialised at the mid point. There are 3 marks on the scale dividing it into 4 quarters to aid workers with internal calibration. They are given no other instructions or guidelines.

	All	Good	Bad
β_1	-0.03^{*}	-0.01	-0.04^{*}
β_2	0.45^{***}	0.66^{***}	0.23^{***}

Table 2: Autocorrelation coefficient β_1 for the AFFECTIVE dataset.

In this paper, we base our analysis on the adequacy dataset of Graham et al. (2015), on Spanish-English newswire data from WMT 2013 (Bojar et al., 2013). The dataset consists of 12 HITS of 100 sentence pairs each; each HIT is annotated by at least 15 workers.

HITs are designed to include quality control items to filter out poor quality scores. In addition to 70 MT system translations, each HIT contains degraded versions of 10 of these translations, 10 reference translations by a human expert corresponding to 10 of these translations, and repeats of another 10 translations. Good workers are assumed to give high scores to the references, similar scores to the pair of repeats, and high scores to the MT system translations when compared to corresponding degraded translations. Workers who submitted scores of clearly bad quality were rejected. For the remaining workers, the Wilcoxon rank-sum test is used to test whether the score difference between the repeat judgements is less than the score difference between translations and the corresponding degraded versions. We divide these workers into "good" and "moderate" based on the threshold of $p < 0.05$.

To eliminate differences due to different internal scales, every individual worker's scores are standardised by subtracting the mean and dividing by the standard deviation of their scores. Following Graham et al. (2015), we use the average of standardised scores of at least 15 good workers as the ground truth.

We refer to the final dataset as "MT$_{adeq}$".

Results As this is a (practically) continuous output, we use a linear regression model, whereby the current score is predicted based on the previous score, with the mean of all worker scores as control. We also controlled for worker correlation with mean score, and position of the sentence in the HIT, but these were not significant and did not affect the autocorrelation. As seen in Table 3, we see a small but significant positive autocorrelation for good workers. The bias is much stronger with

	Good	Moderate	Bad
β_1	0.030^{***}	0.037^{***}	0.192^{***}
β_2	0.741	0.661	0.256
N items	48216	24696	17738

Table 3: MT_{adeq} dataset: Autocorrelation coefficient β_1, showing sequence bias of good, moderate and bad workers.

Position	Good	Moderate	Bad
1st Tertile	0.044^{***}	0.063^{***}	0.179^{***}
2nd Tertile	0.032^{***}	0.034^{***}	0.173^{***}
3rd Tertile	0.015^{**}	0.014^{*}	0.225^{***}

Table 4: MT_{adeq} dataset: Regression coefficient β_1 of adequacy scores with the previous score. We also show results for translations in the first, second or third tertile based on the position of the sentence of the HIT

bad (rejected) workers.

An interesting question is whether the bias changes as workers annotate more data, which could be ascribed to learning through the task, calibrating their internal scales, or becoming fatigued on a monotonous task. Each HIT consists of 100 sentences, and we divide the dataset into 3 equal groups based on the position of sentence in the HIT. As shown in Table 4, for good and moderate workers, the bias is stronger in the first group of sentences annotated, decreases in the second, and is much smaller in the last. This could be because workers are familiarising themselves with the task earlier on, and calibrating their scale. There is no such trend with bad quality scores, possibly because the workers are not putting in sufficient effort to produce accurate scores.

Next we assess the impact of the bias in the worst case situation. We discretize scores into low, middle and high based on equal-frequency binning, and divide the dataset into 3 groups based on the score assigned to the previous sentence. As shown in Table 5 we can see that the sentences in the "low" partition and the "high" partition have a difference of 0.18, which is highly significant;[3] moreover, this difference is likely to be sufficiently large to alter the rankings of systems in an evaluation. The bias remains even when we increase the number of workers and use the average score, as all workers scored the translations in the same order. This shows that the mean is also affected by

[3] $p < 0.001$ using Welch's two-sample t-test

N	All	Low	Middle	High	H − L
1	0.01	−0.09	0.05	0.08	0.18^{***}
5	0.00	−0.05	−0.02	0.08	0.14^{***}
10	−0.00	−0.05	−0.04	0.09	0.13^{***}
15	−0.00	−0.05	−0.02	0.07	0.12^{***}

Table 5: MT_{adeq} dataset: Translations following a low quality translation receive a lower score than those following a good translation: "All" is the mean score of all sentences in the dataset, where each sentence score is calculated as the average of N (standardised) worker scores. "Low", "Middle", and "High" are mean scores of sentences where the previous sentence annotated is of low, medium and high quality, resp. "H − L" is the difference between the average high and low scores.

sequence bias.

Thus, it is theoretically possible to exploit sequence bias to artificially deflate (or inflate) a specific system's computed score by ordering a HIT such that the system's output is seen consistently immediately after a bad (or good) output.

4 Discussion and Conclusions

We have shown significant sequence effects across several independent crowdsourced datasets: a negative autocorrelation in the RTE and TEMPORAL datasets, and a positive autocorrelation in the MT_{adeq} dataset. The negative autocorrelation can be attributed either to sequential contrast effects or the gambler's fallacy. These effects were not significant for the AFFECTIVE dataset, perhaps due to the nature of the annotation task, whereby annotations of one emotion are separated by six other annotations, thus limiting the potential for sequencing effects. It is also possible that the dataset is too small to obtain statistical significance.

MT judgements are subjective, and when people are asked to rate them on a continuous scale, they need time to calibrate their scale. We show that the sequential bias decreases for better workers as they annotate more sentences in the HIT, indicating a learning effect. Since the ordering of the systems is random, system scores obtained by averaging scores of all sentences translated by the system would be unbiased, assuming a sufficiently large sample of sentences. Thus we do not expect sequential bias to have a marked effect on system rankings or other macro-level conclusions on the basis of this data. However, the scores of in-

dividual translations remain biased, which augurs poorly for the use of these annotations at the sentence level, such as when used in error analysis or for training automatic metrics.

Sequence problems can be easily addressed by adequate randomisation — providing each individual worker with a separate dataset that has been randomised, such that no two workers see the same ordered data. In this way sequence bias effects can be considered as independent noise sources, rather than a systematic bias, and consequently the aggregate results over several workers will remain unbiased.

This study has shown that sequence bias is real, and can distort evaluation and annotation exercises with crowd-workers. We limited our scope to binary and continuous responses, however it is likely that sequence effects are prevalent for multinomial and structured outputs, e.g., in discourse and parsing, where priming is known to have a significant effect (Reitter et al., 2006). Another important question for future work is whether sequence bias is detectable in expert annotators, not just crowd workers.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work was supported in part by the Australian Research Council.

References

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles, USA.

Daniel Chen, Tobias J. Moskowitz, and Kelly Shue. 2016. Decision-making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics*.

Lysann Damisch, Thomas Mussweiler, and Henning Plessner. 2006. Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied*, 12(3):166.

R Edward Geiselman, Nancy A Haight, and Lori G Kimata. 1984. Context effects on the perceived physical attractiveness of faces. *Journal of Experimental Social Psychology*, 20(5):409–424.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):330.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, pages 1183–1191, Denver, USA.

Samuel M. Hartzmark and Kelly Shue. to appear. A tough act to follow: Contrast effects in financial markets. *Journal of Finance*.

Tobias Hoßfeld, Matthias Hirth, Judith Redi, Filippo Mazza, Pavel Korshunov, Babak Naderi, Michael Seufert, Bruno Gardlo, Sebastian Egger, and Christian Keimel. 2014. Best practices and recommendations for crowdsourced QoE — lessons learned from the Qualinet Task Force "Crowdsourcing". Lessons learned from the Qualinet Task Force "Crowdsourcing" COST Action IC1003 European Network on Quality of Experience in Multimedia Systems and Services (QUALINET).

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York, USA.

Douglas T. Kenrick and Sara E. Gutierres. 1980. Contrast effects and judgments of physical attractiveness: When beauty becomes a social problem. *Journal of Personality and Social Psychology*, 38(1):131.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121, New York, USA.

J. Rebecca Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association of Computational Linguistics*, 2(1):311–326.

David Reitter, Frank Keller, and Johanna D Moore. 2006. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 121–124, New York, USA.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast — but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, USA.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France.

Timothy D. Wilson, Christopher E. Houston, Kathryn M. Etling, and Nancy Brekke. 1996. A new look at anchoring effects: basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, 125(4):387.