Making Heads *and* Tails of Models with Marginal Calibration for Sparse Tagsets

Michael Kranzlein

Nelson F. Liu

Nathan Schneider

Background
What is calibration?

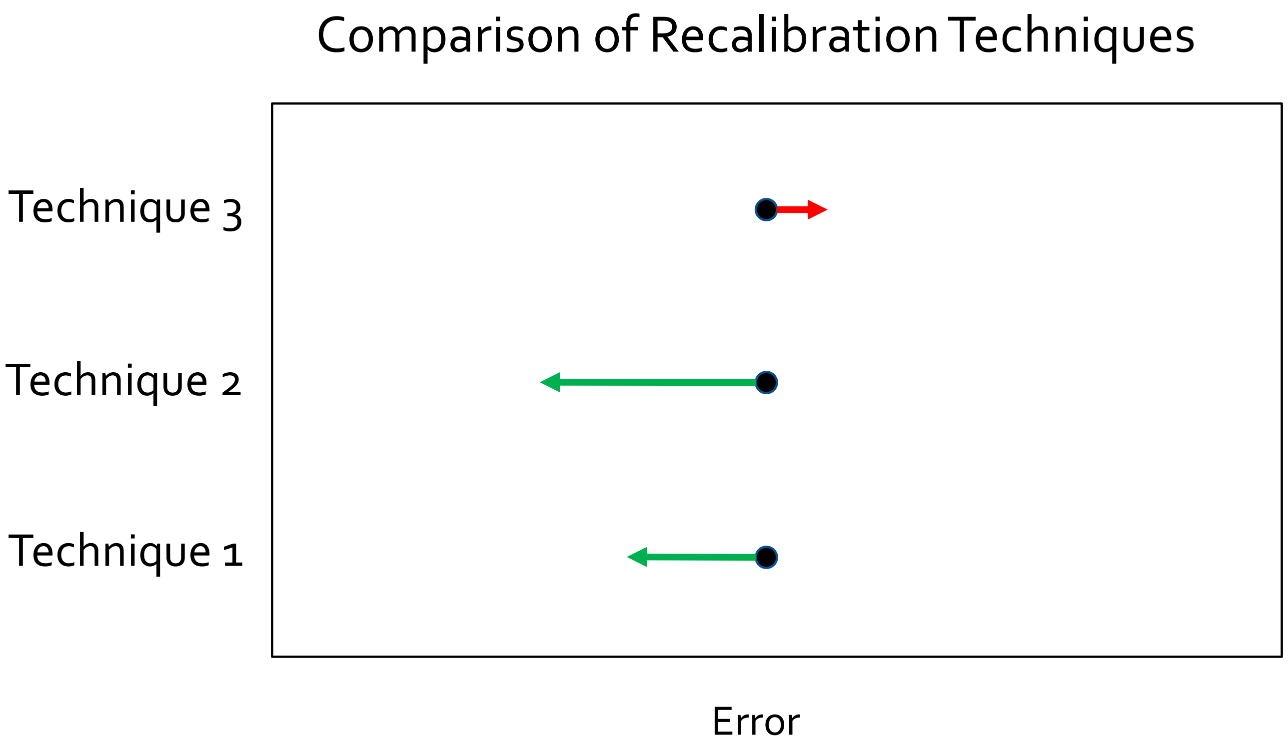
- A model is well calibrated when its probabilities correlate well with empirical accuracy
  - α% of model outputs of probability α should be correct
- A model can be very accurate but also be severely miscalibrated (Guo et al., 2017)
- Reducing calibration error is important
  - Gives you more reliable and interpretable confidence scores
  - Reliable confidence scores may improve results on other tasks or make certain tasks easier
    - Preannotation
    - Rare instance discovery

Background
What is calibration?

- How do we measure calibration error?
  - Ideally, take many sample outputs from the model where the probability is α and see how many are correct
  - Models output continuous scores
    - Suppose $\alpha = 82.53046\%$
    - We probably won't be able to find multiple probabilities α
  - Instead of looking for α exactly, look for similar scores and put them in a *bin*; then calculate deviation from average score and label in the bin
  - Error is an average of the deviations in each bin, weighted by the number of items in each bin

Background
What is calibration?

- We can measure calibration error with uncalibrated scores and recalibrated scores and (hopefully) observe a reduction



Background
What is calibration?

How do we *re*-calibrate a model's probabilities?

1. Incorporate calibration error into objective function during training
2. **Use post-processing techniques that shift scores in a way that minimizes calibration error on held-out data to learn a recalibration model**

Background

Why is it difficult to recalibrate models with sparse tagsets?

- Prior work primarily focuses on top-label calibration
  - Recalibrates only the score for the tag the model predicts for each input
- Sparse tagsets (especially for NLP) are understudied
  - Most existing work is on image classification tasks with balanced, smaller tagsets
- Marginal recalibration typically requires lots of data for each class
  - Ideal approach is developing an independent recalibration model for each class (Kumar et al., 2019)
  - When that's not possible due to lack of data, Shared Classwise Binning (Patel et al., 2021) creates a shared recalibration model among all classes

Methodology

Tag Frequency Grouping (ours)

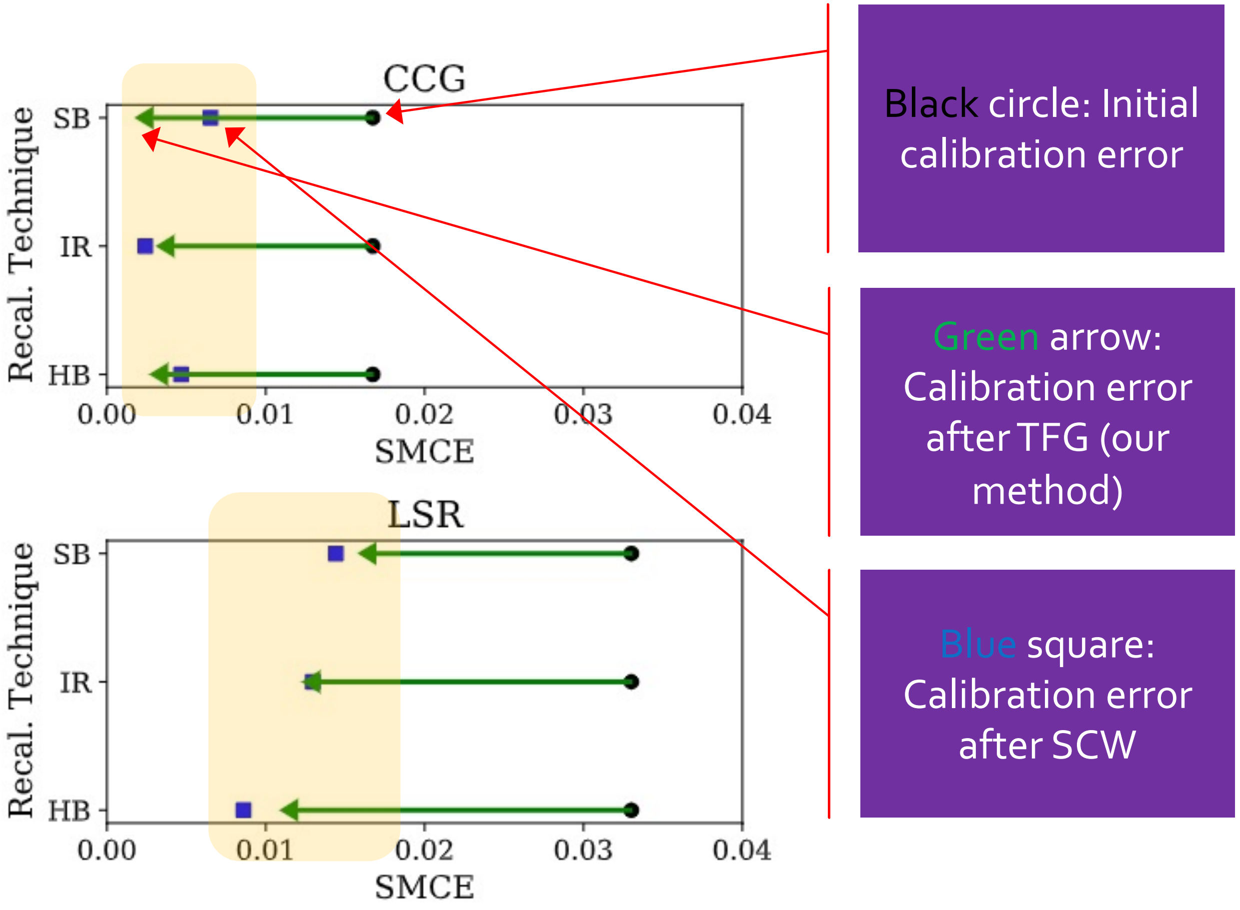
- We hypothesize that tags that are similarly frequent in the training data will be similarly miscalibrated
  - The model may tend to be:
    - Overconfident on the tags it has seen the most
    - Underconfident on rare tags
- Idea: calibrate similarly frequent tags together
  - Sort tags by gold label frequency
  - Divide tags into *G* groups of roughly equal size
  - Calibrate each group together

Experiments

- Compare Shared Classwise Binning (SCW) and Tag Frequency Grouping (TFG) using three techniques on two tasks
- Techniques
  1. Histogram binning (Zadrozny and Elkan, 2001)
  2. Isotonic Regression (Zadrozny and Elkan, 2002)
  3. Scaling Binning (Kumar et al., 2019)
- Tasks
  1. Combinatory Categorial Grammar supertagging (Prange et al., 2021)
  2. Lexical Semantic Recognition (Liu et al., 2021)
- Both tasks have hundreds of tags

Results
(overall)

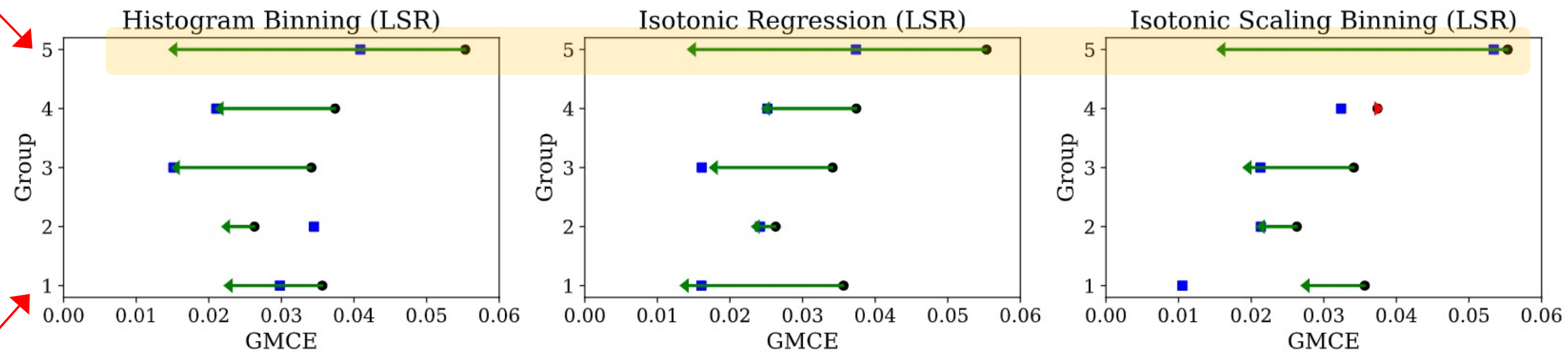
TFG and SCW both do a good job of reducing calibration error on each task!



Results
(by frequency group)

Group 5: Rarest tags

TFG does **better** than SCW on the rarest tags.



Group 1: Most common tags

Conclusions

We showed:

- SCW and TFG can be used for recalibration *and* evaluation (SCW previously only used for recalibration)
- TFG works well, especially for recalibrating scores for rare tags
- TFG in evaluation allows for more fine-grained analysis of calibration error than SCW

Future work:

- We evaluated on 5 frequency groups (G=5); what's the optimal way to determine G?
- CCG and LSR tagsets have structure; can their subtags be used to determine tag groupings?
- Does TFG have benefits for more balanced datasets?

References

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages1321–1330. PMLR.

Ananya Kumar, Percy Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. In Advances in Neural Information Processing Systems, pages 3792–3803. Curran Associates, Inc.

Nelson F. Liu, Daniel Hershcovich, Michael Kranzlein, and Nathan Schneider. 2021. Lexical semantic recognition. In Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021), pages49–56, Online. Association for Computational Linguistics.

Kanil Patel, William H. Beluch, Bin Yang, Michael Pfeiffer, and Dan Zhang. 2021. Multi-class uncertainty calibration via mutual information maximization-based binning. In International Conference on Learning Representations.

Jakob Prange, Nathan Schneider, and Vivek Srikumar. 2021. Supertagging the long tail with tree-structured decoding of complex categories. Transactions of the Association for Computational Linguistics, 9:243–260.

Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In Proceedings of the Eighteenth International Conference on Machine Learning, pages 609–616. Morgan Kaufmann.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the Eighth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, pages 694–699.

Thanks!