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

Michael Kranzlein
Georgetown University
mmk119@georgetown.edu

Nelson F. Liu
Stanford University
nfliu@cs.stanford.edu

Nathan Schneider
Georgetown University
nathan.schneider@georgetown.edu

Abstract

For interpreting the behavior of a probabilistic model, it is useful to measure a model's *calibration*—the extent to which the model produces reliable confidence scores. We address the open problem of calibration for tagging models with *sparse tagsets*, and recommend strategies to measure and reduce calibration error (CE) in such models. We show that several post-hoc *re*calibration techniques all reduce calibration error across the marginal distribution for two existing sequence taggers. Moreover, we propose *tag frequency grouping* (TFG) as a way to measure calibration error in different frequency bands. Further, recalibrating each group separately promotes a more equitable reduction of calibration error across the tag frequency spectrum.

1 Introduction

An advantage of probabilistic models is that, in addition to providing a prediction, they also quantify *uncertainty*. Knowing how certain a model is about a particular prediction can be crucial when using its output for downstream tasks or when weighing its trustworthiness. Of course, the probability estimate associated with a predicted output is an artifact of the model, and is subject to error—separate from the accuracy or error of the prediction itself.

By and large, NLP evaluations of multiclass classifiers and structured prediction models consider only the top *prediction* for an input and how closely it matches the gold standard. Only in some studies is the *probability* assigned to the prediction taken into account at all (e.g. via a precision-recall curve).

A more comprehensive evaluation would examine whether the model's probabilities are *well-calibrated*, i.e., whether they correlate well with empirical accuracy (such that $\approx \alpha\%$ of predictions with probability close to α are in fact correct). Guo et al. (2017) showed that despite high accuracy, modern neural networks can still suffer from severe miscalibration. Fortunately, calibration error is not completely random, and can be corrected post hoc with a second model fit on development data (or even a separate recalibration set if available) as in several *recalibration* techniques (§2).

In domains where NLP models help inform human decision-making (e.g., medicine), having a well-calibrated model is essential. Even in less critical domains, a well-calibrated model has potential to benefit rare instance discovery, pre-annotation, and self-training. In this paper we consider a structured prediction setting of particular relevance in NLP: tagging tasks with sparse tagsets—output spaces with a handful of high-frequency tags and many more rare tags.

Many linguistic phenomena follow power law distributions and thus feature a long tail of individually rare events, which, as we will show, makes it nontrivial to measure calibration error with existing methods, including marginal calibration error (MCE), which requires sufficient samples of each class to produce a reliable estimate (Kumar et al., 2019). We evaluate two English sentence taggers[1] with closed sets of 100s of tags that disambiguate word tokens: a Combinatory Categorial Grammar (CCG) syntactic supertagger with 426 tags (Prange et al., 2021), and a Lexical Semantic Recognition (LSR) tagger with 598 tags (Liu et al., 2021).

Our main contributions are the following:

- We posit that evaluation of calibration should go beyond a model's highest-confidence prediction, extending the arguments of Nixon et al. (2020), with a particular focus on sparse tagsets.
- We propose tag frequency grouping (TFG), a novel technique for evaluating and recalibrating groups of similarly frequent tags in a sparse tagging space.

[1] Data, code, and results are available at https://github.com/nert-nlp/calibration_tfg. Hyperparameters are described in §4.2.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4919–4928
November 7–11, 2021. ©2021 Association for Computational Linguistics

- We introduce two new error metrics based on MCE suitable for tasks where insufficient data is available to apply MCE to all tags.
- We compare TFG and shared class-wise binning (SCW) on two sequence tagging tasks.

2 Background

Calibration studies have two components: a recalibration technique and an evaluation metric. We use similar notation as Kumar et al. (2019) to describe both. That is, we assume a multiclass model $f : \mathcal{X} \to \mathcal{Y}$ that produces a real-valued score $f(X)_k \in [0,1]$ for each class $k \in \mathcal{Y}$. In other words, for any input, the model gives $K = |\mathcal{Y}|$ scores. If these predictions are the output of a softmax function (as is typical for the last layer of neural networks), they will sum to 1 and can be interpreted as uncalibrated confidence scores across the distribution of possible classes or tags. The goal of recalibration is to make these confidence scores more reliable.

2.1 Definition and Measurement

There are several metrics for evaluating calibration error, including maximum calibration error (Naeini et al., 2015), Brier Score (Brier, 1950), calibration error (a term used widely in the literature, but here we refer to definition 2.1 in Kumar et al. (2019)), and expected calibration error (Naeini et al., 2015). We focus on the *marginal calibration error* (Kumar et al., 2019), which is a multiclass extension of CE.[2] MCE uses the l_2-norm to measure, for each class, "the difference between the model's probability and the true probability of that class given the model's output": $MCE(f) =$

$$\sqrt{\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\Big[\big(f(X)_k - \mathbb{P}(Y = k \mid f(X)_k)\big)^2\Big]} \quad (1)$$

This metric is the root mean square error of measurements taken from K binary recalibration models, where $\mathbb{P}$ is the true probability that the class is k given $f(X)_k$, which is the model's predicted probability for class k on input X. But one of the problems we quickly encounter with this definition (and similar measures of calibration error) is that with finite data, we cannot actually measure calibration error, since f outputs values in a continuous range. In practice, this is overcome using binning schemes to estimate $\mathbb{P}(Y = k \mid f(X)_k)$. The range $[0,1]$ is partitioned into bins; each score is placed in the appropriate bin; and error is estimated as the deviation between the average confidence of the bin and the proportion of positive labels in the bin (proportion of positive labels is equivalent to accuracy for top-label calibration).

2.2 Recalibration Techniques

We use three techniques for recalibration: histogram binning (Zadrozny and Elkan, 2001), isotonic regression (Zadrozny and Elkan, 2002), and scaling binning (Kumar et al., 2019). All of these are *post-hoc* techniques—they are applied after the model has been trained. In general, recalibration techniques fit into one of two categories: scaling or binning. Binning techniques quantize the interval of confidence scores and only output a fixed number of unique calibrated scores equal to the number of bins used for recalibration. Scaling techniques output continuous calibrated scores. Scaling techniques are generally better at reducing error, but because their output domain is continuous, the binning techniques used for evaluation are prone to underestimating true calibration error. Kull et al. (2019) showed this with experiments on CIFAR-10 (Krizhevsky, 2009) and ImageNet (Russakovsky et al., 2015).

Histogram Binning. Histogram binning is a popular recalibration technique that is simple and fast. The interval $[0,1]$ is subdivided into B subintervals using the confidence scores from the development set.[3] The bin boundaries can be set such that each bin covers a fixed interval (fixed-width binning), or such that each bin includes the same number of data points (adaptive binning; Nguyen and O'Connor, 2015).

Using the boundaries for these B bins, a confidence score from the test set is calibrated by finding the bin it belongs to and outputting the empirical proportion of positive labels among the development scores in that bin. This definition assumes a binary classification setting, but histogram binning can be extended to a multiclass scenario by building a one-vs.-rest model for each class, by using shared classwise binning (SCW; Patel et al., 2021), or by using TFG, described in §3.3.

Isotonic Regression. Isotonic regression is a scaling technique that fits a non-decreasing piece-

[2] Kull et al. (2019) introduce a metric similar to MCE they call classwise-ECE.

[3] This is also referred to as a recalibration set in the literature, though they need not necessarily be disjoint.

wise linear function on the recalibration set by minimizing the square error subject to the non-decreasing constraint. It produces calibrated scores in a continuous range via linear interpolation.

Scaling Binning. Scaling techniques and binning techniques each have disadvantages. For example, histogram binning usually yields worse results than temperature scaling (another scaling technique), but its error measurement is reliable (Kumar et al., 2019). Scaling binning combines the best of both approaches by first learning a scaling function. Uncalibrated scores are binned, and instead of outputting the proportion of positive labels (as in histogram binning), the calibrated score is the average output of the scaling function on the development scores in the bin. In our experiments with scaling binning, we use isotonic regression as the scaling function.

2.3 Related Work

Zadrozny and Elkan (2002) initially proposed the one-vs.-rest approach for multiclass probabilities. Kuleshov and Liang (2015) recognize the sparsity problem and suggest reducing multiclass calibration of structured prediction to targeted "events of interest" and training a binary forecaster to learn calibrated probabilities of the event happening. This work is extended by Jagannatha and Yu (2020), who treat a sequence of tags as a compositional model output and develop a forecaster based on gradient boosted decision trees. They achieve reductions in expected calibration error and a slight increase in model performance after reranking. Reranking refers to the process of normalizing calibrated scores and reordering them. With most recalibration techniques, it is rare for the ranking to be affected, and with some techniques like isotonic regression, the ranking of calibrated confidence scores will always match the uncalibrated ones.

3 Designing and Evaluating Recalibration Models for Sparse Tagsets

The long tail of tags for CCG and LSR is of particular interest with respect to calibration. Kumar et al. (2019) point out that most studies of multiclass calibration focus primarily on *top-label* calibration (reducing calibration error for only the top prediction out of the model for each input), also called top-1 or top-k when looking at several of the model's top predictions. While *top-label* scores are an important component of calibration, they don't tell the whole story, and we argue that the rest of the distribution (*marginal* calibration) shouldn't be ignored. Recent works that address marginal calibration (Kumar et al., 2019; Patel et al., 2021; Nixon et al., 2020) make similar arguments but still tend to focus on balanced datasets like CIFAR-100, which contains 600 examples for each of 100 classes, or datasets with fewer tags like MNIST (LeCun et al., 1998), MNIST Fashion (Xiao et al., 2017), and CIFAR-10, which each have 10 classes.

In our analysis of marginal calibration, we study two long tails of distributions related to calibrating a sparse tagset: low confidence scores and low-frequency tags. We show how the standard *one-vs.-rest* approach to multiclass calibration becomes infeasible as the size of the tagging space grows, and we provide specific recommendations for quantifying calibration error with sparse tagsets, where the lack of instances of rare tags poses unique challenges.

Extending section 4 of Nixon et al. (2020) with a particular focus on sparse tagsets, we now discuss the many design decisions that need to be made regarding multiclass calibration.

3.1 Thresholding

While we are interested in calibrating more of the distribution than is addressed with top-label calibration, it would be unwise to include all confidence scores. This is more an issue for evaluation than for recalibration. The justification for this decision is made clear in the distribution of the confidence scores and in prior work (Nixon et al., 2020). We observe that more than 98% of our two models' (which each have hundreds of possible tags) confidence scores are below 0.0001. Evaluating a recalibration model on all scores is likely to underestimate the error of the model, where the error on more likely output candidates will be washed out by excessively many near-zero scores that often have little error (particularly on a highly accurate model).

Instead, we select a threshold t and if any scores are below this threshold, they are excluded from the recalibration and evaluation sets. For isotonic regression, including the scores below t would have little effect as this scaling technique produces a piecewise function independent of any hyperparameter for the number of recalibration bins required for other techniques. However, if a threshold is not applied with binning techniques, many bins

will contain only near-zero scores. For this reason (and consistency), we apply the threshold t before both recalibration and evaluation for all techniques. Consequently, in our results we report calibration error on unnormalized scores, since thresholding excludes data and prevents us from obtaining calibrated scores for all tags in the distribution.

3.2 Binning

How should bin boundaries be determined? With a sparse tagset, it is even more important to avoid fixed-width binning, especially as the number of bins increases. Fixed-width binning will lead to significant imbalance, whereby the bins covering intervals of lowest and highest confidence scores will have many more items per bin, and the bins in the middle of the range will have very few items, causing high variance in estimates of calibration error. Thresholding does make the distribution less skewed, but many of the confidence scores in both of our datasets are low even after a threshold is applied. The alternative to fixed-width binning, adaptive binning (Nguyen and O'Connor, 2015), puts the same number of items in each bin, leading to wider bins in the middle of the range, but guarantees each bin will have a sufficient number of data points for recalibration to overcome sampling error.

How to avoid too-small bins due to rare tags? Marginal calibration error as defined in eq. (1) treats each class as a binary recalibration problem and averages the error in each recalibration model that was estimated by binning. Nixon et al. (2020) highlight that a finer-grained, per-class approach to evaluation analagous to MCE is ideal because it allows "systematic differences in the calibration error between classes to be evaluated without washing each other out." In contrast to MCE, the top-label approach measures error only among the model's highest confidence score for each input (i.e. the confidence score associated with the model's predicted label). This is done by binarizing the multiclass problem via one-hot labels. The top prediction of the model is selected and its gold label is taken to be 1 if that class is the true class and 0 otherwise. In this way, confidence scores for multiple tags can be evaluated together. This idea is key to how we modify MCE to evaluate our recalibration models.

While MCE is the gold standard, it requires ample data in all tags in order to get a reliable measurement. With our sparse tagsets, measuring MCE separately for each tag is unfortunately infeasible, since we would not have enough samples in each bin. Nguyen and O'Connor (2015), for instance, recommend 200 samples per bin to reduce sampling error. In the literature, the floor for the number of bins used in evaluation is around 5. Assuming 5 bins at ≥200 samples each, that means creating a tag-specific recalibration model would require 1000 confidence scores.

On its face, this is not a huge ask for marginal calibration with no thresholding, since having a recalibration set of 1000 tokens will produce 1000 confidence scores for each tag. But the number of near-zero confidence scores will increase as the tagset grows, and these near-zero scores are not as relevant to a discussion about calibration as actual candidate outputs from the model. For top-label calibration, it is possible to build a strong recalibration model, but in order to measure MCE for that model (with our assumption of 5 bins and at least 200 scores per bin), we would need at least 1000 tokens *where each tag is predicted*. So the relative frequency of the rarest tag controls the total number of instances required for reliable binning (e.g., a tag occurring at a rate of 1% would necessitate a recalibration set of 100,000 instances).

We experiment with two strategies to overcome this and derive a modified MCE metric. First, we extend the binarization approach of top-label error measurement to all labels, effectively creating a shared binning model for collective evaluation. This approach, shared classwise binning (SCW), was introduced by Patel et al. (2021) for recalibration, but is extendable to evaluation. (We will introduce TFG, a generalization of SCW giving finer control over the sharing, in §3.3.)

For SCW evaluation, we modify MCE and introduce shared marginal calibration error (SMCE). When operationalized with binning, we get eq. (2). $\mathcal{D}$ contains the set of above-threshold confidence scores for all tokens and tags in the data. In this equation, $\overline{q}_b$ is the average confidence score of the b-th bin and $\overline{p}_b$ is the average of the binary labels associated with each confidence score in the b-th bin. N is the total number of confidence scores being recalibrated. $AdaBin(\mathcal{D}, \beta)$ is our adaptive binning function that partitions the sorted confidence scores into bins of size β. A key difference between this metric and MCE is that scores for multiple tags are included in the square.

$$SMCE(\mathcal{D}, \beta) = \sqrt{\sum_{b \in AdaBin(\mathcal{D}, \beta)} \frac{|b|}{N} (\overline{q}_b - \overline{p}_b)^2} \quad (2)$$

Using SCW for recalibration simply means learning a single recalibration model, pooling together all confidence scores from all tags.

How many bins should we use? We report results using 10 bins for recalibration and evaluation in our experiments, to ensure each bin has a sufficient number of datapoints.

3.3 Tag Frequency Grouping

As we have explained, SCW solves the problem of rare tags by pooling all tags together when recalibrating or evaluating calibration error. But a concern is that this may be too coarse-grained: all tags are not necessarily created equal with respect to their calibration. We therefore propose a new technique, TFG, to strike a balance between the two extremes of treating all tags together or independently with respect to calibration. TFG, like SCW, can be used for recalibration, evaluation, or both.

The intuition is simple. We often find that models are overconfident with tags seen frequently in the training data and underconfident with tags seen less frequently. Therefore, we hypothesize that **tags that are similarly frequent in the training data will be miscalibrated in similar ways**, and that by grouping together tags of similar frequencies and developing a separate recalibration model for each group, we can achieve improved results over SCW and calibrate tags that lack sufficient data for a class-specific recalibration model. The number of groups G should be selected such that $G \ll K$, and in this paper, we report results where $G = 5$.[4]

Choosing an optimal value of G is tricky. As G increases, the amount of recalibration data available for each group decreases, making each recalibration model less reliable. However, too low a value can lead to a reduced benefit over SCW with the loss of granularity (both in the recalibration models and in evaluation). Higher values for G are likely suitable for larger datasets that still suffer from sparsity. However, if the dataset is sufficiently large and balanced, we recommend that independent recalibration models be created for each tag instead of using TFG or SCW.

[4] Patel et al. (2021) explored a similar idea in one of their experiments on digit recognition: digits with similar class priors were grouped together manually for recalibration. However, Patel et al. did not propose a general grouping technique, nor did they address large sparse tagsets as we do here.

In order to maximize generalization, we propose constructing tag groups based not on a model's output, but on the gold tag frequencies in the training data. The procedure is simple—sort the tags by descending frequency, and add the next most frequent tag to the group until the number of instances with gold tags in that group is greater than or equal to $1/G$.

Figure 1 depicts a hypothetical example of TFG on a training set with 45 instances. Note that there's some overflow in the first group. This overflow can occur in any group except the last one, and in theory could lead to a worst-case scenario where the last group is much smaller than the others. In practice, this is unlikely to occur, but making sure all tag groups encompass a similar amount of training data is a good step to take prior to recalibration.

SMCE (eq. (2)) can be adapted to grouped marginal calibration error (GMCE) for TFG by replacing $\mathcal{D}$, which contains confidence scores for all tags, with $\mathcal{G} \subseteq \mathcal{D}$, which contains confidence scores for one group (a subset of tags):

$$GMCE(\mathcal{G}, \beta) = \sqrt{\sum_{b \in AdaBin(\mathcal{G}, \beta)} \frac{|b|}{N} (\overline{q}_b - \overline{p}_b)^2} \quad (3)$$

4 Experiments

In our experiments, we develop recalibration models for two taggers with sparse tagsets and measure the improvement over the uncalibrated confidence scores with SMCE (overall error) and GMCE (per-group error).

4.1 Taggers

We consider two supervised tagging tasks trained and evaluated on different English datasets: CCG supertagging—a syntactic task with a large amount of training data and a high-accuracy model, and Lexical Semantic Recognition—a semantic task with less data and a lower-accuracy model.

4.1.1 CCG Supertagging

CCG is a lexicalized grammar formalism that is frequently used for syntactic and semantic parsing. CCG supertagging is the task of labeling each token with a complex, structured label that belies its function (Clark, 2002; Bangalore and Joshi, 2010). Bangalore and Joshi (1999) describe supertagging as "almost parsing", because a sequence of supertags maps a sentence to a small set of possible parses—the CCGBank (Hockenmaier and Steedman, 2007)

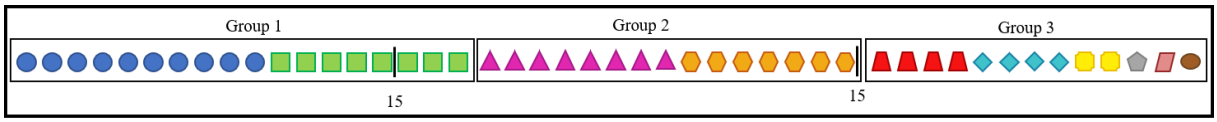

Figure 1: Illustration of tag frequency grouping (TFG) with 45 training instances and $G = 3$ tag groups. Each shape represents a gold tag from the training data. Tags are sorted by frequency. Starting with the most frequent tag, groups are formed by iteratively adding all instances of a tag until the size of the group equals or exceeds the number of training instances divided by the number of groups. When TFG is used for recalibration (as opposed to just evaluation), a separate recalibration model is learned for each group.

dataset has over 1,200 unique CCG labels. By convention, the model is limited to predicting only tags that appeared at least 10 times in the training data, yielding 425 tags + the UNK tag. We use the non-constructive BERT-based (Devlin et al., 2019) model from (Prange et al., 2021) with its default hyperparameters. The tagger was trained on 927,497 tokens and obtained a dev accuracy of 96.1%.

4.1.2 Lexical Semantic Recognition

LSR involves joint identification of multi-word expressions (MWEs), classification of lexical units, and disambiguation of coarse-grained supersenses and for noun, verb, preposition, and possessive expressions (Liu et al., 2021). Liu et al. (2021) model this task as a sequence labeling problem using the STREUSLE dataset (Schneider and Smith, 2015; Schneider et al., 2018). For each token, they predict a tag with the conjunction of the token's MWE, lexcat, and supersense. Their model is also based on BERT, but it uses a conditional random field (CRF; Lafferty et al., 2001) for decoding. We use a version of the model with no training or decoding constraints that has 598 tags and use its default hyperparameters. The tagger was trained on 44,801 tokens and obtained a dev accuracy of 81.1%. To extract marginal distributions from the CRF, we use the Forward-Backward algorithm.

4.2 Experimental Overview

We use three techniques for recalibration: histogram binning, isotonic regression, and scaling binning with SCW and TFG. We use standard splits from the LSR and CCG datasets, fitting recalibration models on the development set and evaluating on the test set.

We exclude the one-vs.-rest recalibration setup from our experiments. The infeasibility of this approach with sparse tagsets is in fact one of the motivations for this paper. With SCW, there is one recalibration model per technique, and with TFG, there are G independent recalibration models. We do not normalize the calibrated scores, since thresholding excludes many tags from the distribution on each sample.

For both grouping approaches and all three techniques, we evaluate tags in their respective frequency groups (GMCE) and collectively (SMCE). Evaluating with GMCE gives us more insight into which tags are miscalibrated (both before and after recalibration) and reduces exposure to cancellation effects among the different tags that could lead to an underestimation of error. Recall that an average of independent per-tag evaluations is the gold standard for mitigating these effects but is not possible due to how many tags lack sufficient representation in our datasets.

Summary of Hyperparameters and Reacalibration Model Design Following our explanations from §3, we made the following decisions for our models:

- Apply a threshold and exclude all model predictions less than .01
- Use adaptive binning with 10 bins
- Use the l_2 norm for evaluation
- Evaluate error on unnormalized scores
- Set $G = 5$ for recalibration and evaluation with TFG

5 Results and Discussion

Our experimental results are visually summarized in figures 2 and 3. Table 1 provides total error across the marginal distribution as well as the error in the most frequent tags and least frequent tags.

Overall, our models for both datasets benefit from recalibration and see substantial reductions in calibration error with SCW, TFG, and all recalibration techniques. Relative to the CCG model, the LSR model has higher absolute error, and we see greater relative improvements from recalibration. The recalibrated CCG model has the lowest *absolute* error.

How do post-hoc techniques compare? We evaluated three recalibration techniques in our experiments: histogram binning, isotonic regression,

		CCG						LSR					
Recalibration		**All**		**Group 1**		**Group 5**		**All**		**Group 1**		**Group 5**	
Method	G	SMCE	Δ	GMCE	Δ	GMCE	Δ	SMCE	Δ	GMCE	Δ	GMCE	Δ
None	—	*.0167*		*.0183*		*.0396*		*.0330*		*.0356*		*.0553*	
Scaling Binning	1	.0065	−61.0%	.0225	22.89%	.0390	−1.47%	.0144	−56.36%	.0105	−70.48%	.0534	−3.48%
Scaling Binning	5	.0019	−88.87%	.0049	−73.09%	.0114	−71.19%	.0159	−51.83%	.0269	−24.42%	.0153	−72.27%
Isotonic Regres.	1	.0024	−85.54%	.0226	23.72%	.0276	−30.35%	.0130	−60.74%	.0161	−54.95%	.0374	−32.49%
Isotonic Regres.	5	.0032	−80.93%	.0230	25.58%	.0228	−42.37%	.0124	−62.57%	.0132	−62.93%	.0142	−74.34%
Histogram Bin.	1	.0047	−72.06%	.0286	56.49%	.0422	6.54%	.0086	−73.94%	.0298	−16.4%	.0409	−26.15%
Histogram Bin.	5	.0028	−83.42%	.0254	38.69%	.0218	−44.93%	.0110	−66.76%	.0222	−37.78%	.0145	−73.71%
N		72,373		12,873		15,854		15,933		2,854		3,020	
Tag types		415		1		382		377		3		302	
Tag freq in train				[22.2%,22.2%]		[.0%,0.4%]				[7.1%,10.3%]		[.0%,.1%]	
Tokens		55,371		12,873		9,167		5,381		2,739		1,716	

Table 1: Marginal calibration error (measured with SMCE and GMCE) before and after recalibration with different techniques on two tasks: Combinatory Categorial Grammar (CCG) supertagging and Lexical Semantic Recognition (LSR). These data are visualized in figure 3. SMCE indicates shared marginal calibration error, and GMCE indicates grouped marginal calibration error (see §3.2); Δ refers to the relative change over the original model (lower is better). 5 groups are used for tag frequency–based evaluation; only the highest-frequency tags (Group 1) and lowest-frequency tags (Group 5) are shown. The TFG conditions use the same 5 groups for separate recalibration models, while the SCW conditions ($G = 1$) use multiple groups only for evaluation.

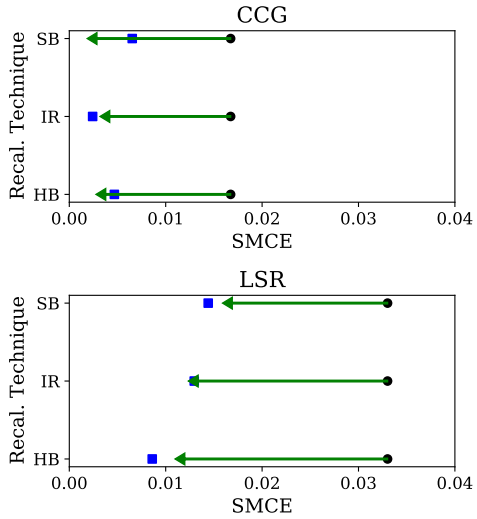

Figure 2: Evaluation of recalibration techniques, TFG, and SCW using SMCE. Techniques include histogram binning (HB), isotonic regression (IR), and scaling binning (SB) using isotonic regression as the scaling function. Black circles show initial calibration error; green arrows pointing to the left show reductions in calibration error after recalibration with TFG ($G = 5$); and blue squares show calibration error after recalibration with SCW. Lower SMCE is better.

and scaling binning. When measuring calibration error collectively in figure 2, we noticed similar performance. Breaking the error down by tag group in figure 3 gives us more insights about how our recalibration techniques affect tags of different frequency. All of the techniques achieve similar performance, though isotonic regression has the fewest outliers, with only one situation—Group 1 for CCG—where calibration error gets worse.

Both binning techniques, and in particular histogram binning, are susceptible to making things worse in some cases. This happens more with the CCG tagger, which was fairly well calibrated to begin with. It is more accurate than the LSR tagger and has high average confidence in its output, with relatively few confidence scores near 50%.

In recalibration with binning methods, this makes CCG more prone to unlucky wide bin boundaries (which are more likely to have high error). Using more bins for recalibration could help mitigate this problem; we used 10 bins for both models for parity in comparisons. While isotonic regression appears the most reliable, it does not have the same quantifiable error bounds as scaling binning (Kumar et al., 2019), which should be taken into account when choosing a recalibration technique.

How do groups compare? For both datasets, Group 2 has the lowest initial calibration error, and it sees some of the smallest changes after recalibration. These tags are still frequent in the training data, but less so than the tags in Group 1. Group 5, which contains the rarest tags, has the highest calibration error and sees the biggest improvements.

The statistics at the bottom of table 1 show us how unbalanced our tagsets truly are. N shows the number of confidence scores that exceed the threshold. Then there is the number of tags represented in each group and the minimum and maximum training frequencies of the tags in each group. "Tokens" shows the number of tokens with any score above the threshold in each group. Remarkably, Group

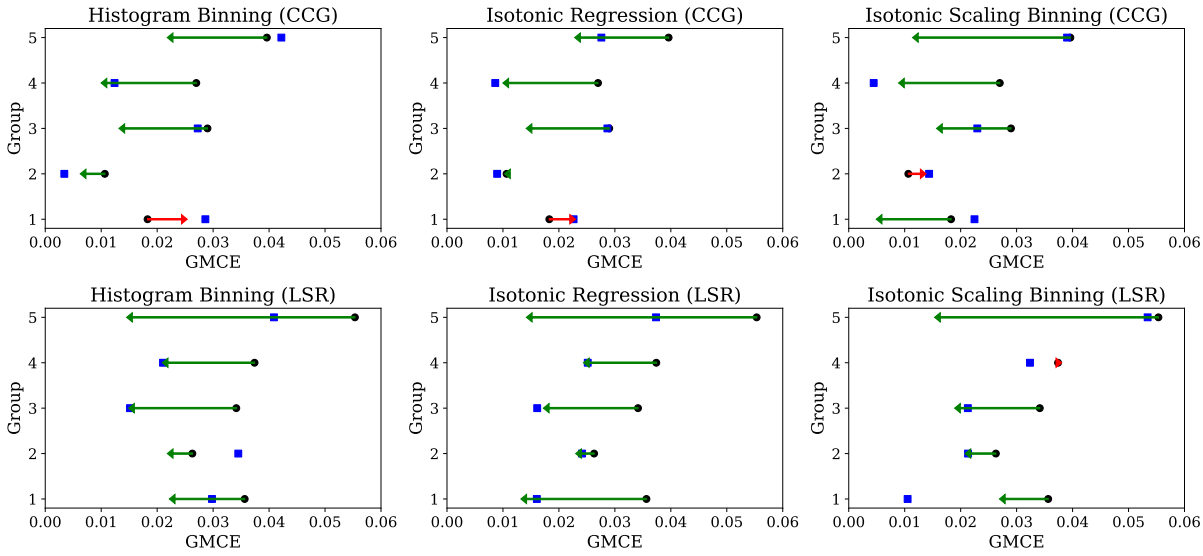

Figure 3: Evaluation of recalibration techniques, TFG, and SCW on 5 groups of tags. Colors and shapes are the same as figure 2, but red arrows pointing to the right indicate an increase in calibration error. Group 1 includes the most frequent tags (as represented in the training data), and Group 5 includes the rarest tags. Lower GMCE is better.

1 for CCG has just 1 tag, `N`, that makes up more than 22% of the gold-labeled data in the training set, whereas the most frequent tag for Group 5 is just .4% of the gold-labeled data.

How does TFG compare to SCW for recalibration? TFG performs drastically better than SCW on the rarest tags. In most cases for other tag groups, the TFG results and SCW results are close. Only with scaling binning on LSR for Group 1 does SCW outperform TFG by a wide margin. This may be the result of a lucky bin boundary, as SCW does worse than TFG with histogram binning for LSR.

Groups 1 and 2 are interesting for CCG. With SCW, all three techniques *increased* calibration error for Group 1. With TFG, histogram binning appears to sacrifice performance on Group 1 for the benefit of Group 2, and scaling binning does the opposite.

TFG yields strong improvements in all other tag groups for CCG, whereas SCW does not. The only other case where TFG slightly increases calibration error is Group 4 for LSR with scaling binning.

Our results suggest that when used for recalibration, TFG yields overall improvements in calibration error that are similar to or better than SCW, especially on less frequent tags. For datasets where SCW might outperform TFG, we can still recommend TFG for evaluation of models with sparse tagsets via GMCE, since GMCE provides more information about which tags suffer from the greatest miscalibration.

6 Conclusion

We examined the challenges of evaluating and reducing calibration error with sparse tagsets. In particular, we introduced TFG to offer more control over how tags are pooled together given that some are too infrequent to be recalibrated/evaluated independently. We showed that SCW and TFG are easily extensible from recalibration to the evaluation setting with the SMCE and GMCE metrics, and that GMCE gives more specific insight into where in a tag distribution the most calibration error exists and where it can be reduced. On one semantic task and one syntactic task, we found substantial improvement in calibration error for the head and tail of the tag distribution.

Opportunities for further research include devising methods for choosing and evaluating the optimal value for G and considering normalizing scores despite the elimination of scores below the threshold. While the recalibrated model would be unable to assign any confidence to tags excluded by thresholding, this effect may be minimal, and it could lead to improved interpretability since the distribution would sum to 1.

It may also be worth relying not just on frequency but incorporating the structure of each tag into the grouping process. LSR and CCG tags, for example, are compositional, and could be grouped based on subtag. Testing whether TFG has benefits for more balanced tagsets is another opportunity.

Acknowledgements

We are grateful to anonymous reviewers as well as members of the NERT lab for their feedback on this work. This research was supported in part by NSF award IIS-1812778 and grant 2016375 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel.

References

Srinivas Bangalore and Aravind K. Joshi. 1999. Supertagging: an approach to almost parsing. *Computational Linguistics*, 25(2):237–265.

Srinivas Bangalore and Aravind K Joshi. 2010. *Supertagging: Using Complex Lexical Descriptions in Natural Language Processing*. The MIT Press.

Glenn W Brier. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1).

Stephen Clark. 2002. Supertagging for Combinatory Categorial Grammar. In *Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+6)*, pages 19–24, Universitá di Venezia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, pages 1321–1330. PMLR.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Abhyuday Jagannatha and Hong Yu. 2020. Calibrating structured output predictors for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2078–2092, Online. Association for Computational Linguistics.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. Technical report.

Volodymyr Kuleshov and Percy Liang. 2015. Calibrated structured prediction. In *Advances in Neural Information Processing Systems 28*, Montréal, Canada.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *Advances in Neural Information Processing Systems 32*, pages 12316–12326, Vancouver, Canada.

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

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yann LeCun, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

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)*, pages 49–56, Online. Association for Computational Linguistics.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2901–2907, Austin, Texas.

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1587–1598, Lisbon, Portugal. Association for Computational Linguistics.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. 2020. Measuring calibration in deep learning. *arXiv:1904.01685 [cs, stat]*.

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.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196, Melbourne, Australia. Association for Computational Linguistics.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547, Denver, Colorado. Association for Computational Linguistics.

Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747 [cs, stat]*.

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 Eigth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699.