Is the Best Better? Bayesian Statistical Model Comparison for Natural Language Processing

Piotr Szymański
Department of Computational Intelligence
Wrocław University of Science and Technology
piotr.szymanski@pwr.edu.pl

Kyle Gorman
Graduate Program in Linguistics
City University of New York
kgorman@gc.cuny.edu

Abstract

Recent work raises concerns about the use of standard splits to compare natural language processing models. We propose a Bayesian statistical model comparison technique which uses k-fold cross-validation across multiple data sets to estimate the likelihood that one model will outperform the other, or that the two will produce practically equivalent results. We use this technique to rank six English part-of-speech taggers across two data sets and three evaluation metrics.

1 Introduction

Gorman and Bedrick (2019) raise concerns about standard procedures used to compare speech and language processing models. They evaluate the performance of six English part-of-speech taggers using multiple randomly-generated training-testing splits; in some cases, they fail to reproduce previously-published system rankings established using a single "standard" split. They argue that point estimates of performance derived from a single training-testing splits are insufficient to establish system rankings, even when null hypothesis significance testing is used for model comparison.

In this study, we propose a technique for system comparison based on Bayesian statistical analysis. Our approach, motivated in Section 2 and described in Section 3, allow us to infer the likelihood that one model will outperform the other, or even that both models' performance will be practically equivalent, something that is not possible with the frequentist statistical tests used by Gorman and Bedrick. Our approach can also be applied simultaneously across multiple data sets. As a proof of concept, in Sections 4–5 we apply the proposed method using the experimental setup of Gorman and Bedrick, and in Section 6, we use it to rank the six taggers, compare evaluation metrics, and interpret the results. Our failure to reproduce some of earlier reported results leads us to discuss the impact of repeating experiments, contrasting performance in multiple measures and the advantages of comparing likelihoods in Section 6. We also discuss the notion of *practical equivalence* for speech and language technology.

2 Prior work

Langley (1988) argues that machine learning should be viewed as an experimental science, and as such, machine learning technologies should be evaluated according to their performance on multiple held-out data sets. Dietterich (1998) proposes a framework for comparing two supervised classifiers using a null hypothesis tests to determine whether two classifiers have the same likelihood of predicting a correct result. This study introduces several methods, including a paired t-test for k-fold cross-validation results. However, he notes that the assumptions of normality and independence may not be satisfied in all cases. Nadeau and Bengio (2000) propose a correlation-based correction for the Dietterich t-test procedure which adjusts for the overlap between folds. Hull (1994) and Schütze et al. (1995) propose non-parametric tests for comparing models across multiple data sets; Salzberg (1997) proposes Bonferroni-corrected ANOVA analysis. Demšar (2006) reports that the Friedman non-parametric test with the Nemenyi correction makes fewer assumptions and has greater power than parametric tests. Other authors (e.g., Luengo et al., 2009; García et al., 2010; Derrac et al., 2011) further adapt the Friedman test for model comparison.

However, as Demšar (2006) notes, there still does not exist a non-parametric null hypothesis test designed for use with a repeated measure (i.e., k-fold) design across multiple data sets. As a result,

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2203–2212,
November 16–20, 2020.

there is no procedure that takes into consideration the variance in scores of a given data set, at least within the frequentist paradigm. Demšar (2008) and Benavoli et al. (2017) enumerate additional problems with null hypothesis significance testing (NHST) procedures for model comparison:

- NHST does not estimate probabilities for hypotheses; i.e., it does not tell us how likely two models are to perfrom equivalently,
- NHST p-values conflate effect size and sample size; i.e., with a sufficiently large sample, one can claim significance even if the effect size is trivial,
- NHST yields no information about the null hypothesis; i.e., one cannot draw further conclusions from a failure to reject the null hypothesis, and
- there is no principled way to select an appropriate α-level for NHST.

These issues lead Benavoli et al. to reject NHST-based model comparison in favor of a Bayesian approach. Bayesian hypothesis tests are defined by a likelihood function $p(d \mid \theta)$, a probability model of the data d conditioned on θ, a vector of parameters. The prior distribution for θ, $p(\theta)$ must also be defined. From these components, a posterior probability distribution $p(\theta \mid d)$ can then be calculated and queried (i.e., sampled from) to perform inference. Various techniques can be used to estimate θ; they are usually related to the differences in models' scores using some evaluation metric, and ultimately, to whether one method is likely to perform better or worse than the other. Thus, the posterior distribution can be used to perform model comparison. Benavoli et al. also introduce the notion of a *region of practical equivalence* (henceforth, ROPE), which allows Bayesian hypothesis testing to estimate the likelihood that two models' results will be functionally indistinguishable. ROPE defines an interval around a model's result - if another model's performance falls within this interval - they are deemed practically equivalent. For example, if one deems that a difference of 1 percentage point in accuracy between models denotes practical equivalence, a $[-0.01, 0.01]$ interval is used as ROPE. If one model performs at .941 accuracy and another at .949 - they will be deemed practically equivalent. This allows to protect the statistical procedure from artifacts and false alarms of significance. Readers are referred to the accessible tutorial by Benavoli et al. (2017) for further details.

Corani et al. (2017) generalize Bayesian model comparison to a repeated measures scenario in which there are multiple data sets with unequal score variances. They propose a hierarchical Bayesian model for estimating the likelihood of one model performing better, worse, or equivalently, to another. We now proceed to briefly describe and adapt this procedure to re-evaluate the findings of Gorman and Bedrick (2019).

3 Bayesian model comparison

Imagine a scenario where one wishes to compare the performance of two classifiers across q data sets. By performing m k-fold evaluations, the experimenter obtains a vector of $n = mk$ observations, i.e., differences in scores, between the two models: $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,n})$. The values in these vectors are a positive cross-correlation ϱ because cross-validation introduces overlap in training data. Let δ_i denote the mean difference score on the ith data set, and let δ_0 denote the average population-level difference. Corani et al. (2017) propose a hierarchical probabilistic model

$$\begin{aligned} \mathbf{x}_i &\sim \text{MVN}(1\delta_i, \Sigma_i), \\ \delta_1 \ldots \delta_q &\sim t(\mu_0, \sigma_0, \nu), \\ \sigma_1 \ldots \sigma_q &\sim \text{unif}(0, \bar{\sigma}) \end{aligned} \tag{1}$$

where MVN is a multivariate normal distribution over the vector of classifier differences with mean δ_i and a covariance matrix Σ_i with variance σ_i^2 along the diagonal and $\varrho\sigma_i^2$ on the off-diagonal. Data set variances are drawn from a Student distribution parameterized by the average population-level difference δ_0 and variance σ_0, with μ degrees of freedom. The prior distributions for δ_0, σ_0, and μ are defined so as to preserve the robustness of the model; these are motivated and described in more detail by Corani et al. (2017). Crucially, we model the differences obtained in individual runs using a multi-variate normal distribution oriented to the per-data set mean differences with a per-data set variance, and the mean differences using a unimodal distribution robust to outliers and non-normality. Per-data set variances are modeled by a uniform distribution.

After the model learns the parameter distributions from experimental data, we obtain a posterior probability distribution $p(\delta_0, \sigma_0, \mu \mid d)$. To infer whether one classifier is more likely to outperform another—or whether they are practically equivalent—we draw N_s samples from the posterior distribution. We use decision counters n_{left}, n_{rope}, and n_{right} to keep track of how many times the left model was more likely to outperform the right model, to be practically equivalent to the right model, and to be outperformed by the right model, respectively. For each sample of the parameters, we define the posterior of the mean difference accuracy on a new unseen data set δ_{next} as $t(\delta_0, \sigma_0, \mu)$. We obtain the outcome probabilities by integrating the distribution over the three intervals—e.g., we obtain the probability that the left model is better than the right by integrating from the left end of the distribution to the left edge of the ROPE interval, and so on—and then incrementing the decision counter for the region with the highest outcome probability. Finally, we compute likelihoods for the three scenarios by dividing the decision counts by the number of samples drawn: $P(left) = \frac{n_{left}}{N_s}$, $P(rope) = \frac{n_{rope}}{N_s}$, and $P(right) = \frac{n_{right}}{N_s}$.

Instead of significance, we thus estimate the likelihoods that one method is better than the other (or are practically equivalent). These estimates follow from observing the beliefs of a Bayesian model that models the probability of the methods' mean difference on unseen data sets, after sampling parameters from a meta-distribution which estimates the difference and variance over the population of data sets with μ degrees of freedom.

4 Materials and methods

To compare the results of our study with the ones obtained by Gorman and Bedrick (2019), we use the same models, data sets, and evaluation metrics.[1] That is, we compare implementations of the TnT (Brants, 2000), Collins (Collins, 2002), LAPOS (Tsuruoka et al., 2011), Stanford (Manning, 2011), NLP4J (Choi, 2016), and Flair (Akbik et al., 2018) part-of-speech taggers using the Wall St. Journal portions of the Penn Treebank (v. 3; Marcus et al., 1993) and OntoNotes (v. 5; Weischedel et al., 2011), two widely-used corpora of American English financial news. Summary statistics for this data are given in Table 1.

[1] http://github.com/kylebgorman/SOTA-taggers

	# sentences	# tokens
Penn Treebank	49,208	1,173,766
OntoNotes	37,025	901,673

Table 1: Summary statistics for the two corpora.

We perform 20 randomized 10-fold cross validations, obtaining 200 measurements of each model's performance on each data set. In each run, 80% of the data is used for training, 10% for validation, and 10% for evaluation. We fit Bayesian models using the `baycomp` library[2] and draw 50,000 samples from the posterior.

Following Gorman and Bedrick, we use three evaluation metrics. Token accuracy is simply the number of test data tokens correctly tagged divided by the total number of tokens, and is the standard intrinsic evaluation metric used for this task. OOV accuracy is similar to token accuracy but is restricted to out-of-vocabulary tokens, i.e., those found in the test data but not in the training data. Finally, sentence accuracy is the number of test data sentences that contain no tagging errors, divided by the number of test sentences. Ground-truth data is provided by human annotators.[3]

5 Results

Posterior distributions of the hierarchical models are visualized in Figures 1–3 and summarized in Table 2. We define the ROPE to have the same size as the 95% confidence interval; this is roughly 2–3% for sentence and OOV accuracy, and 0.2% for token accuracy. Thus, two models are judged to be practically equivalent in sentence accuracy if they differ in performance on fewer than 98 sentences of the Penn Treebank or 75 sentences on the slightly smaller OntoNotes corpus. For token accuracy, they are practically equivalent if they differ on fewer than 210 PTB tokens or 162 OntoNotes tokens, respectively.

The hierarchical model estimates, for example, that TnT, the simplest tagger, would be outperformed in token accuracy by any of the other five taggers 80-90% of the time. However, there is a surprisingly high chance of practical equivalence in token accuracy between the Collins tagger, LAPOS, and the Stanford tagger; for instance,

[2] http://github.com/janezd/baycomp

[3] Annotation quality for these data has been studied by Ratnaparkhi (1997) and Manning (2011), among others.

		Token accuracy			Sentence accuracy			OOV accuracy		
		$>$	$\approx$	$<$	$>$	$\approx$	$<$	$>$	$\approx$	$<$
TnT	Collins	.168	.003	**.829**	.125	.001	**.874**	.158	**.550**	.292
	LAPOS	.137	.001	**.862**	.127	.000	**.873**	.162	.305	**.533**
	Stanford	.109	.000	**.891**	.112	.000	**.888**	.088	.010	**.903**
	NLP4J	.152	.000	**.848**	.156	.000	**.844**	.116	.002	**.882**
	Flair	.158	.000	**.842**	.136	.000	**.864**	.094	.000	**.906**
Collins	LAPOS	.116	**.617**	.267	.105	.215	**.680**	.063	**.842**	.095
	Stanford	.180	**.441**	.379	.124	.120	**.756**	.118	.038	**.845**
	NLP4J	.137	.014	**.848**	.153	.010	**.837**	.166	.010	**.824**
	Flair	.164	.000	**.836**	.157	.000	**.843**	.138	.001	**.861**
LAPOS	Stanford	.099	**.822**	.079	.084	**.829**	.087	.138	.091	**.771**
	NLP4J	.172	.112	**.716**	.163	.137	**.700**	.161	.018	**.821**
	Flair	.192	.004	**.805**	.190	.001	**.809**	.127	.003	**.870**
Stanford	NLP4J	.206	.122	**.672**	.191	.200	**.609**	.148	**.441**	.411
	Flair	.197	.001	**.802**	.190	.001	**.809**	.130	.058	**.812**
NLP4J	Flair	.150	.055	**.795**	.148	.024	**.827**	.092	**.619**	.288

Table 2: Token, sentence, and OOV accuracy ranking likelihoods.

the probability of practical equivalence of the latter two is 84%. This result is contrary to Gorman and Bedrick's replication of a standard split evaluation—they report that LAPOS is significantly better than the Collins tagger, and that the Stanford tagger is significantly better than LAPOS, according to two-tailed McNemar tests at $\alpha = .05$—but it is consistent with their subsequent failure to consistently reproduce this ranking using randomly-generated splits and Bonferroni-corrected McNemar tests. In contrast, NLP4J and Flair are quite likely to outperform the other taggers, and Flair has an 80% chance of outperforming NLP4J.

Similar results are obtained with sentence accuracy, a less-commonly used metric. TnT is once again quite likely to be outperformed by other models. Whereas LAPOS is quite likely to outperform the Collins tagger, there is an 82% probability that LAPOS and Stanford taggers will yield practically equivalent results. Both NLP4J and Flair are both quite likely to outperform earlier models, and Flair is most likely to outperform NLP4J.

There is a 55% chance of practical equivalence between TnT and the Collins tagger for OOV accuracy. This is somewhat surprising because the two models use rather different strategies for OOV inference: TnT estimates hidden Markov model emission probabilities for OOVs using a simple suffix-based heuristic (Brants, 2000, 225f.), whereas the Collins tagger, a discriminatively-trained model, uses sub-word features developed by Ratnaparkhi (1997) to handle rare or unseen words. Similarly, whereas NLP4J and Flair also use distinct OOV modeling strategies, we estimate that they have a 62% likelihood to achieve practical equivalence on this metric.

6 Discussion

Using the methods above, we obtain the following likelihood-based performance rankings:

- token accuracy: TnT $<$ Collins $\approx$ LAPOS $\approx$ Stanford $<$ NLP4J $<$ Flair,
- sentence accuracy: TnT $<$ Collins $<$ LAPOS $\approx$ Stanford $<$ NLP4J $<$ Flair, and
- OOV accuracy: TnT $\approx$ Collins $\approx$ LAPOS $<$ Stanford $\lessapprox$ NLP4J $\approx$ Flair.

We also find some divergences from the results reported by Gorman and Bedrick. For instance, they find that the Stanford tagger has significantly higher token accuracy than LAPOS on the Penn Treebank standard split. According to our model, the two taggers are most likely practically equivalent, a result which is consistent with their later finding that Stanford outperforms LAPOS on only 1 out of 20 Penn Treebank random splits. We also find out that while both taggers were practically

equivalent in both token and sentence accuracy, Stanford is likely to outperform LAPOS in OOV words, which could have impacted the statistical significance in the original experiment, as the repetition of the k-fold procedure causes strong variation - of the vocabulary available at training and OOV token sets - between experimental runs.

We note that Bayesian comparison and the precise quantities it estimate may give insights into the particular strengths and weaknesses of the various models and evaluation metrics. For instance, we infer that whereas the Collins tagger improves upon TnT, and Flair improves upon NLP4J, in both token and sentence accuracy, these improvements are not likely to be due to differences in the models' handling of out-of-vocabulary words. This is because TnT and the Collins tagger, and NLP4J and Flair, are most likely practically equivalent in their tagging accuracy for OOV words.

In Bayesian approaches as we are thinking about probabilities of a method outperforming another method. As a result we can do what was not possible in the NHST approach taken by Gorman and Bedrick. We can order methods into at a partial ordering to gain an insight into which methods are more likely to perform better than others. We can do this based on the modeled likelihoods, but it would not be in a NHST framework, because there are currently no multiple comparison correction procedures that take into account the variance of repeated runs of a method on the same data set.

Gorman and Bedrick reported that LAPOS would be sure to outperform Collins on PTB (20 out of 20 times), but not on Ontonotes (7 out of 20 times) in token accuracy. We found out that that the most likely scenario, when the performance is modeled using a hierarchical model on evidence from both data sets jointly, that these difference are likely within practical equivalence.

We set the interval of practical equivalence of observed accuracies to match the 95% confidence intervals reported by Bedrick and Gorman, to maintain a capacity for comparing the two experimental approaches. However, we believe it is much more useful to have an interpretable and intuitively understandable definition of what practical equivalence means in the experiment. Instead of setting it based on statistical confidence intervals, we recommend selecting the ROPE to represent the scale of human annotator differences, or the error level that does not negatively impact a downstream task that depends on the prediction quality of evaluated methods.

7 Conclusions

We compare the performance of six part-of-speech taggers on two data sets using twenty repetitions of a ten-fold cross-validation procedure and statistical system comparison performed using hierarchical Bayesian models. By sampling from the posterior distribution of these models, we estimate the likelihood that a given tagger will be better than, worse than, or practically equivalent to other taggers on three different evaluation metrics. These estimates are valid insofar as the data sets used to estimate the Bayesian models comprise a representative sample of a coherent population of data sets. This method provides a principled way to perform statistical model comparison using k-fold cross-validation, a data-efficient evaluation technique. It also allows us to incorporate results obtained across multiple data sets and to make population-level inferences. We finally compare the results obtained with the proposed method to those computed using randomly generated splits and traditional NHST-based model comparison. The results provide new insights into the strengths and weaknesses of English part-of-speech tagging models, complementing other approaches to model comparison and interpretation.

Acknowledgments

We would like to thank Steve Bedrick for previous work on this topic. This work was supported by the statutory funds of the Department of Computational Intelligence, Wroclaw University of Science and Technology.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embedding for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. 2017. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18(1):2653–2688.

Thorsten Brants. 2000. TnT: a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231.

Jinho D. Choi. 2016. Dynamic feature induction: the last gist to the state-of-the-art. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 271–281.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 1–8.

Giorgio Corani, Alessio Benavoli, Janez Demšar, Francesca Mangili, and Marco Zaffalon. 2017. Statistical comparison of classifiers through Bayesian hierarchical modelling. *Machine Learning*, 106(11):1817–1837.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Janez Demšar. 2008. On the appropriateness of statistical tests in machine learning. In *Workshop on Evaluation Methods for Machine Learning*.

Joaquín Derrac, Salvador García, Daniel Molina, and Francisco Herrera. 2011. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1):3–18.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924.

Salvador García, Alberto Fernández, Julián Luengo, and Francisco Herrera. 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180(10):2044–2064.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791.

David A. Hull. 1994. *Information retrieval using statistical classification*. Ph.D. thesis, Stanford University.

Pat Langley. 1988. Machine learning as an experimental science. *Machine Learning*, 3(1):5–8.

Julián Luengo, Salvador García, and Francisco Herrera. 2009. A study on the use of statistical tests for experimentation with neural networks: analysis of parametric test conditions and non-parametric tests. *Expert Systems with Applications*, 36(4):7798–7808.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *12th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 171–189.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Claude Nadeau and Yoshua Bengio. 2000. Inference for the generalization error. In *Advances in Neural Information Processing Systems*, pages 307–313.

Adwait Ratnaparkhi. 1997. A maximum entropy model for part-of-speech tagging. In *Conference on Empirical Methods in Natural Language Processing*, pages 133–142.

Steven L. Salzberg. 1997. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328.

Hinrich Schütze, David A. Hull, and Jan O. Pedersen. 1995. A comparison of classifiers and document representations for the routing problem. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 229–237.

Yoshimasa Tsuruoka, Yusuke Miyao, and Jun'ichi Kazama. 2011. Learning with lookahead: can history-based models rival globally optimized models? In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 238–246.

Ralph Weischedel, Eduard Hovy, Mitchell P. Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: a large training corpus for enhanced processing. In Joseph Olive, Caitlin Christianson, and John McCarthy, editors, *Handbook of natural language processing and machine translation*, pages 54–63. Springer.

A Visualizations

Visualizations of the posterior samples in Section 5 are shown in Figures 1–3 below.

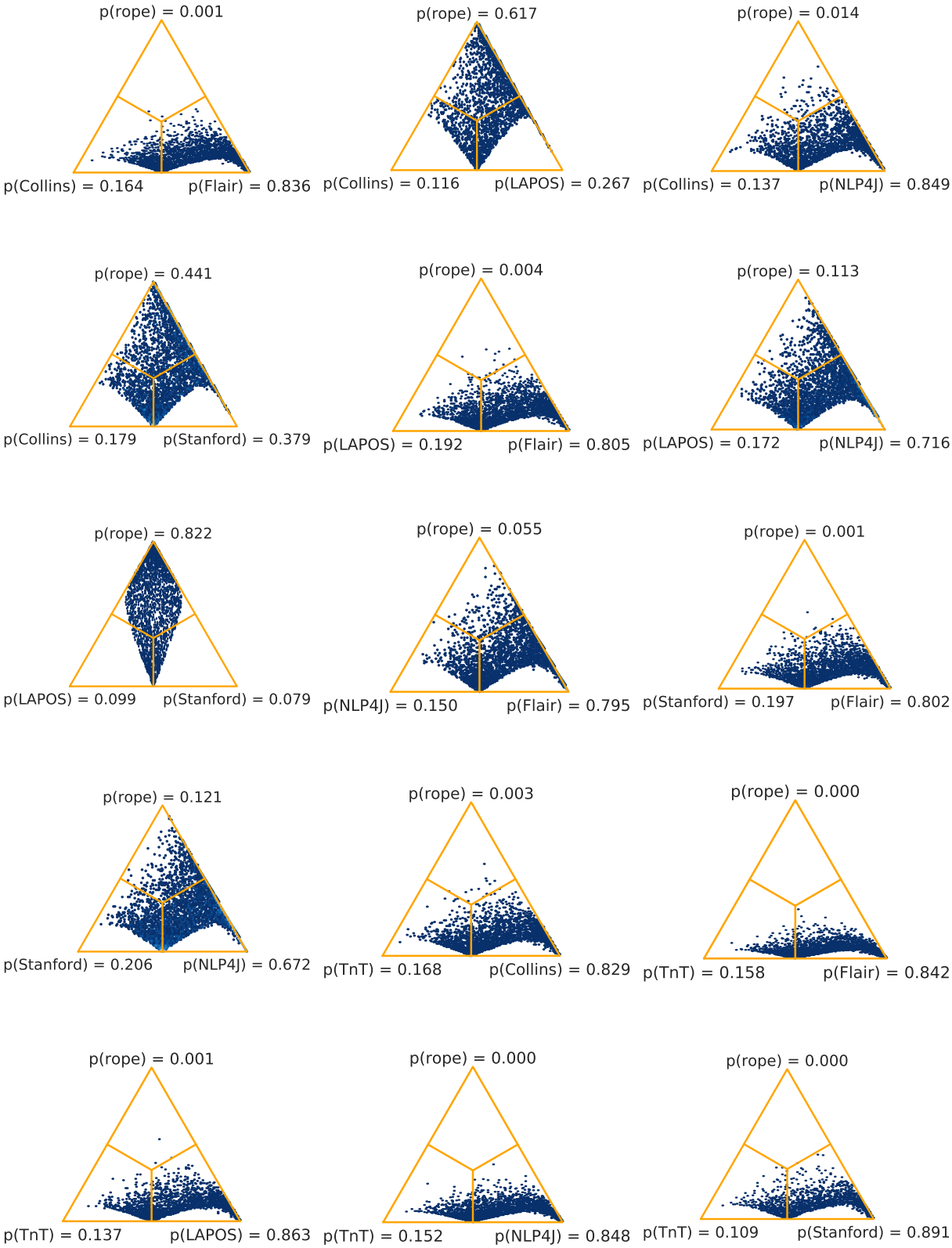

Figure 1: Pairwise comparisons of models' token accuracy; the triangles illustrate 50,000 samples drawn from the posterior distribution, and the likelihood that a given method would perform better, or that their results would be practically equivalent.

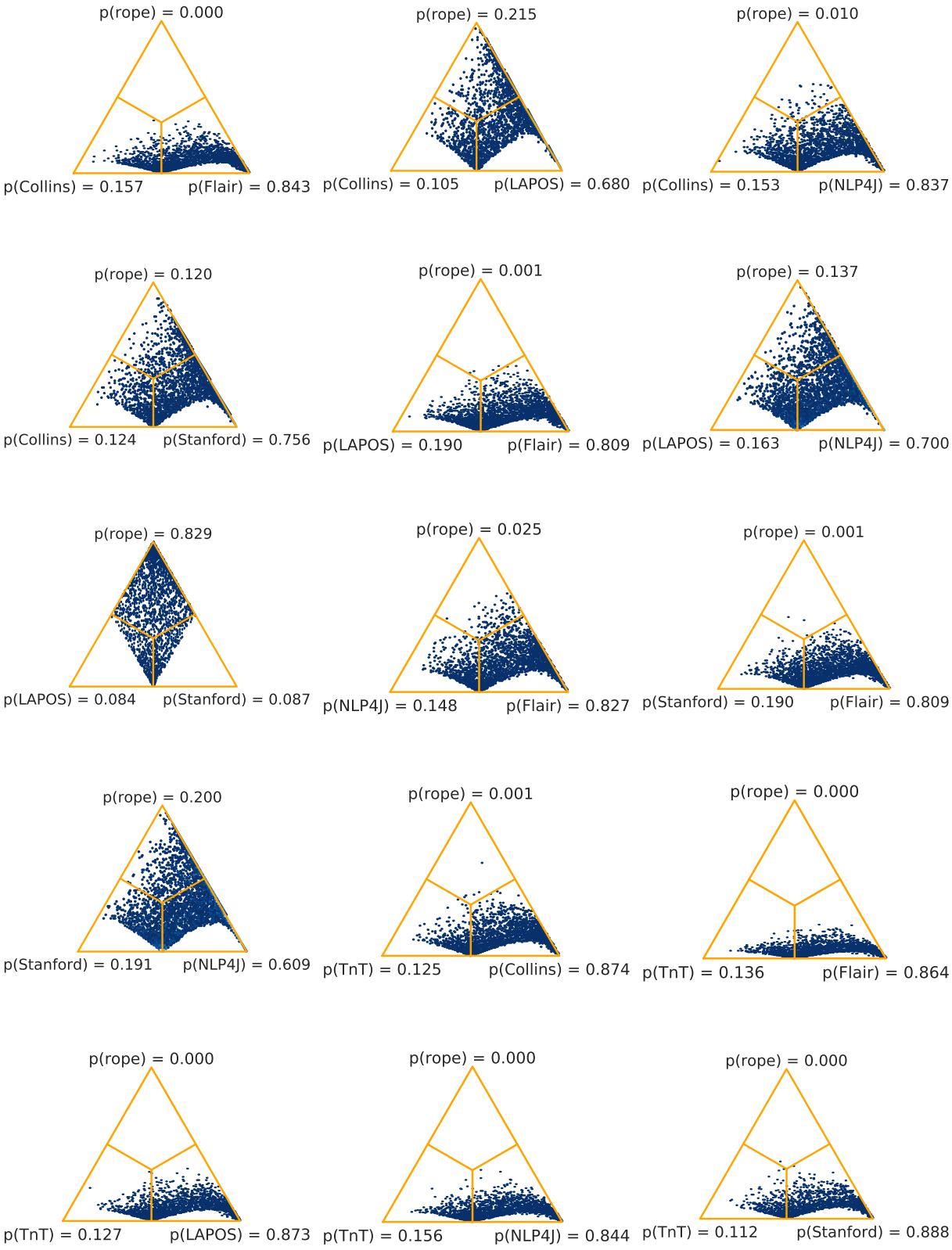

Figure 2: Pairwise comparisons of models' sentence accuracy; the triangles illustrate 50,000 samples drawn from the posterior distribution alongside the likelihood that a given method would perform better, or that their results would be practically equivalent.

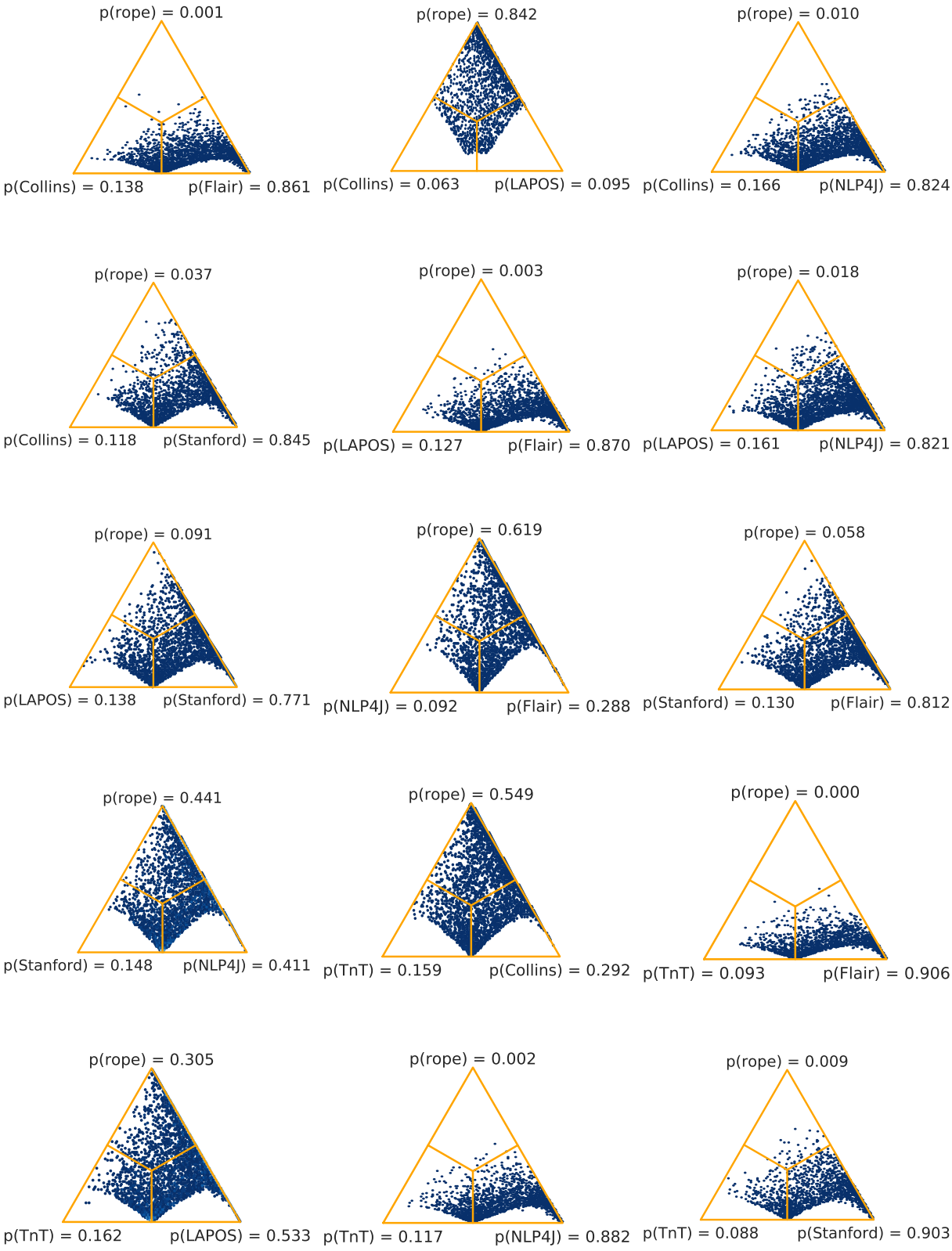

Figure 3: Pairwise comparisons of models' OOV accuracy; the triangles illustrate 50,000 samples drawn from the posterior distribution alongside the likelihood that a given method would perform better, or that their results would be practically equivalent.