Bayes Test of Precision, Recall, and F_1 Measure for Comparison of Two Natural Language Processing Models

Ruibo Wang
School of Modern Education Technology
Shanxi University
Taiyuan, China, 030006
wangruibo@sxu.edu.cn

Jihong Li
School of Software
Shanxi University
Taiyuan, China, 030006
li_ml@sxu.edu.cn

Abstract

Direct comparison on point estimation of the precision (P), recall (R), and F_1 measure of two natural language processing (NLP) models on a common test corpus is unreasonable and results in less replicable conclusions due to a lack of a statistical test. However, the existing t-tests in cross-validation (CV) for model comparison are inappropriate because the distributions of P, R, F_1 are skewed and an interval estimation of P, R, and F_1 based on a t-test may exceed [0,1]. In this study, we propose to use a block-regularized 3×2 CV (3×2 BCV) in model comparison because it could regularize the difference in certain frequency distributions over linguistic units between training and validation sets and yield stable estimators of P, R, and F_1. On the basis of the 3×2 BCV, we calibrate the posterior distributions of P, R, and F_1 and derive an accurate interval estimation of P, R, and F_1. Furthermore, we formulate the comparison into a hypothesis testing problem and propose a novel Bayes test. The test could directly compute the probabilities of the hypotheses on the basis of the posterior distributions and provide more informative decisions than the existing significance t-tests. Three experiments with regard to NLP chunking tasks are conducted, and the results illustrate the validity of the Bayes test.

1 Introduction

The comparison of two models is a key step in natural language processing (NLP) with the precision (P), recall (R), and F_1 measures. The comparison could be described as follows: For two NLP models on a given text corpus, which model produces a higher performance system with a relatively high probability? The direct comparison with a point estimation of P, R, and F_1 on a test corpus is unscientific from a statistical perspective and usually leads to less replicable results (Dror et al., 2017). In reality, the comparison generally could be formalized with a statistical hypothesis testing, and many prominent tests, such as K-fold cross-validated (CV) t-test (Daelemans and Hoste, 2002), 5×2 CV t-test and F-test (Dietterich, 1998; Alpaydin, 1999), and block-regularized 3×2 CV (3×2 BCV) t-test (Wang et al., 2014), have been conducted. However, the distributions of P, R, and F_1 are skewed (Wang et al., 2015) and take values in [0, 1], but an interval estimation of P, R, and F_1 based on a t-test may exceed [0,1].

In this study, we introduce a Bayes test that is more informative than the previous prominent null hypothesis significance testing (NHST) methods in NLP (Dror et al., 2018). The test consists of three main components: (1) a 3×2 BCV (Li et al., 2009; Wang et al., 2014) that provides an optimal partition of corpus and three repetitions of two-fold CV; (2) calibrated posterior distributions and accurate credible intervals (CIs) of P, R, and F_1 instead of a normal approximation; and (3) a Bayes test of P, R, and F_1 that provides the probability of which model outperforms the other.

When partitioning the corpus, certain frequency distributions over linguistic units of the training set should be consistent with that of the validation set. Therefore, partitioning a corpus into two equal parts and conducting a two-fold CV are reasonable for model comparison. In fact, a 3×2 BCV is a specific version of an $m \times 2$ BCV (Wang et al., 2017a) that possesses three repetitions of two-fold CV. The three repetitions are regularized with certain conditions, such as the frequency distribution of the named entity types in a named entity recognition (NER) task, to reduce the unintentional introduced difference in the frequency distributions between the training and validation sets due to the random partitioning of a corpus and to make the comparison more reliable. Particularly, the $m \times 2$ BCV estimator of certain evaluation metrics pos-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4135–4145
Florence, Italy, July 28 - August 2, 2019.

sesses a minimum variance, which ensures that the tests on the 3×2 BCV have higher powers and replicabilities (Wang et al., 2014, 2017b).

Actually, a t distribution is inappropriate for P, R, and F_1 (Yeh, 2000). Wang et al. (2015) have obtained a posterior distribution and a CI of F_1 in a 3×2 BCV, but the distribution did not consider the correlations in the 3×2 BCV estimators, which makes the distribution inaccurate and improper in the comparison.

In this study, accurate posterior distributions and CIs of P, R, and F_1 on the 3×2 BCV are obtained, and a Bayes test is introduced to compare two NLP models. The Bayes test provides the probabilities of the hypotheses in the comparison, which is more informative and reasonable than the conventional NHST. Finally, three experiments in NLP chunking tasks are used to show the validity of the Bayes test.

2 3×2 BCV Posterior Distributions of P, R, and F_1 of an NLP Model

Assume D_n is a text corpus, where n is the count of labeled instances in D_n. For example, n is the count of sentences in an NER corpus.

When computing the P, R, and F_1 of an NLP model, D_n is usually divided into two parts with a partition (S, T) in a hold-out (HO) validation, containing a training set S, a validation set T, and $D_n = S \cup T$. Assume their sizes are $|S| = |T| = n/2$. The confusion matrix on T is $\mathcal{M} = (\text{TP}, \text{FP}, \text{FN}, \text{TN})$, where TP, FP, FN, and TN stand for true positive, false positive, false negative, and true negative, respectively. From these counts, one can compute the P, R and F_1:

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \text{R} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \text{F}_1 = \frac{2\text{PR}}{\text{P} + \text{R}}. \quad (1)$$

Goutte and Gaussier (2005) provided the natural probabilistic interpretations of P and R. Specifically, $\mathcal{M}$ follows a multinomial distribution with parameters $\pi = (\pi_{TP}, \pi_{FP}, \pi_{FN}, \pi_{TN})$ such that $\pi_{TP} + \pi_{FP} + \pi_{FN} + \pi_{TN} = 1$. Then, P and R estimate the following probabilities:

$$p = P(l = + | z = +), \quad r = P(z = + | l = +), \quad (2)$$

where l and z represent the true and predicted labels and + indicates a positive label. Correspondingly, F_1 estimates $f_1 = 2pr/(p + r)$.

Let n_+ denote the count of positive observations in D_n. Let (S, T) be a partition in 3×2 BCV, and the count of positive observations in T satisfies

$$\text{TP} + \text{FN} = n_+/2. \quad (3)$$

2.1 Posterior Distributions of P, R, and F_1 in an HO Validation

Property 2 in (Goutte and Gaussier, 2005) shows that TP|TP + FN follows a binomial distribution with parameters of $n_+/2$ and r. Then,

$$\text{Var}[\text{R}] = \text{Var}\left[\frac{2}{n_+}\text{TP}\right] = \frac{2r(1-r)}{n_+}, \quad (4)$$

where $\text{Var}[\cdot]$ is obtained over D_n. The proof of Eq. (4) is given in the supplemental material.

Assume r follows a beta prior distribution, that is, $r \sim Be(\lambda, \lambda)$, and the posterior distribution of r is $r|\mathcal{M} \sim Be(\text{TP} + \lambda, \text{FN} + \lambda)$ (Goutte and Gaussier, 2005). When $\lambda = 1$, $P(r|\mathcal{M})$ has a mode:

$$mode[r|\mathcal{M}] = \text{R}. \quad (5)$$

Similarly, assume $p \sim Be(\lambda, \lambda)$, and $p|\mathcal{M} \sim Be(\text{TP} + \lambda, \text{FP} + \lambda)$, and its mode is

$$mode[p|\mathcal{M}] = \text{P}. \quad (6)$$

On the basis of the posterior distributions of P and R, Wang et al. (2015) proved that the posterior distribution of F_1 is

$$P(f_1 = t|\mathcal{M}) = \frac{2^a(1-t)^{a-1}(2-t)^{-a-b}t^{b-1}}{B(a, b)}, \quad (7)$$

where $B(\cdot, \cdot)$ is a beta function with parameters $a = \text{FP} + \text{FN} + 2\lambda$ and $b = \text{TP} + \lambda$.

2.2 3×2 BCV

Let $\mathbb{P} = \{(S_j, T_j)\}_{j=1}^3$ denote a partition set of a 3×2 BCV with regularization conditions of $|S_j| = |T_j|$ and $|S_j \cap S_{j'}| \approx n/4$ for $j \neq j'$. Each partition (S_j, T_j) corresponds to a two-fold CV. $\mathbb{P}$ can be constructed with two steps: (a) Divide a text corpus D_n into four equal-sized sub-blocks, denoted as $B_i, i = 1, 2, 3, 4$. (b) Take two sub-blocks as a training set in turn and the other two as a validation set. Table 1 shows the partition set $\mathbb{P}$.

2.3 3×2 BCV Posterior Distributions of P, R, and F_1

Let $\boldsymbol{\mathcal{M}} = \{\mathcal{M}^{(j)}\}_{j=1}^3 = \{(\mathcal{M}_1^{(j)}, \mathcal{M}_2^{(j)})\}_{j=1}^3$ be a collection of confusion matrices in a 3×2 BCV, where confusion matrix $\mathcal{M}_1^{(j)}$ employs the

Index	$\mathbb{P}$	S_j	T_j
1	(S_1, T_1)	B_1,B_2	B_3,B_4
2	(S_2, T_2)	B_1,B_3	B_2,B_4
3	(S_3, T_3)	B_2,B_3	B_1,B_4

Table 1: Partition set of 3×2 BCV.

training set S_j and the validation set T_j in the j-th two-fold CV, and $\mathcal{M}_2^{(j)}$ uses T_j as the training set and S_j as the validation set. Let $\mathcal{M}_k^{(j)} = (\mathrm{TP}_k^{(j)}, \mathrm{FP}_k^{(j)}, \mathrm{FN}_k^{(j)}, \mathrm{TN}_k^{(j)})$.

Here, we aim to infer the posterior distributions $P(p|\boldsymbol{\mathcal{M}})$, $P(r|\boldsymbol{\mathcal{M}})$, and $P(f_1|\boldsymbol{\mathcal{M}})$.

Conditioned on $\boldsymbol{\mathcal{M}}$, the micro-averaged values of P, R, and F_1 are

$$\mathrm{P}_{3\times 2} = \frac{\frac{1}{3}\sum_{j=1}^{3}\frac{1}{2}\sum_{k=1}^{2}\mathrm{TP}_k^{(j)}}{\frac{1}{3}\sum_{j=1}^{3}\frac{1}{2}\sum_{k=1}^{2}(\mathrm{TP}_k^{(j)}+\mathrm{FP}_k^{(j)})} \quad (8)$$

$$\mathrm{R}_{3\times 2} = \frac{\frac{1}{3}\sum_{j=1}^{3}\frac{1}{2}\sum_{k=1}^{2}\mathrm{TP}_k^{(j)}}{\frac{1}{3}\sum_{j=1}^{3}\frac{1}{2}\sum_{k=1}^{2}(\mathrm{TP}_k^{(j)}+\mathrm{FN}_k^{(j)})} \quad (9)$$

$$\mathrm{F}_{1,3\times 2} = \frac{2\mathrm{P}_{3\times 2}\mathrm{R}_{3\times 2}}{\mathrm{P}_{3\times 2}+\mathrm{R}_{3\times 2}}. \quad (10)$$

We first investigate the posterior distribution of R, $P(r|\boldsymbol{\mathcal{M}})$. Considering that $\mathrm{TP}_k^{(j)} + \mathrm{FN}_k^{(j)} = n_+/2$ is a constant (Eq. (3)) unrelated to j and k, $\mathrm{R}_{3\times 2}$ is rewritten as

$$\mathrm{R}_{3\times 2} = \frac{1}{3}\sum_{j=1}^{3}\mathrm{R}^{(j)} = \frac{1}{6}\sum_{j=1}^{3}\sum_{k=1}^{2}\mathrm{R}_k^{(j)}, \quad (11)$$

where

$$\mathrm{R}_k^{(j)} = \frac{\mathrm{TP}_k^{(j)}}{\mathrm{TP}_k^{(j)}+\mathrm{FN}_k^{(j)}}, \quad (12)$$

$$\mathrm{R}^{(j)} = \frac{\frac{1}{2}\sum_{k=1}^{2}\mathrm{TP}_k^{(j)}}{\frac{1}{2}\sum_{k=1}^{2}(\mathrm{TP}_k^{(j)}+\mathrm{FN}_k^{(j)})}. \quad (13)$$

Thus, the variance of $\mathrm{R}_{3\times 2}$ is

$$\mathrm{Var}\left[\mathrm{R}_{3\times 2}\right] = \frac{1+\rho_1+4\rho_2}{3n_+}r(1-r). \quad (14)$$

The proof of Eq. (14) is given in Appendix A. ρ_1 and ρ_2 are two correlation coefficients between the point HO estimators in $\mathrm{R}_{3\times 2}$. The definitions of ρ_1 and ρ_2 are as follows:

- Define $\sigma = \mathrm{Var}\left[\mathrm{R}_k^{(j)}\right]$. According to Eq. (4), we obtain $\sigma = 2r(1-r)/n_+$.
- $\rho_1 = \mathrm{Cov}\left[\mathrm{R}_1^{(j)}, \mathrm{R}_2^{(j)}\right]/\sigma$ is the correlation of two HO estimators in $\mathrm{R}^{(j)}$ in a two-fold CV.
- $\rho_2 = \mathrm{Cov}\left[\mathrm{R}_k^{(j)}, \mathrm{R}_{k'}^{(j')}\right]/\sigma$ is the correlation of two HO estimators of R in different two-fold CVs, where $j \neq j'$ and $k, k' = 1, 2$.

However, the six confusion matrices in $\boldsymbol{\mathcal{M}}$ are correlated because the three partitions are performed on a single text corpus and the training sets contain overlapping samples. Therefore, the likelihood $p(\boldsymbol{\mathcal{M}}|r) \neq \prod_{j=1}^{3}\prod_{k=1}^{2}p(\mathcal{M}_k^{(j)}|r)$. The correlation prevents us to derive a closed form of $p(r|\boldsymbol{\mathcal{M}})$, which is the main challenge in this study.

To overcome the challenge, an **effective confusion matrix** $\mathcal{M}_e = (\mathrm{TP}_e, \mathrm{FP}_e, \mathrm{FN}_e, \mathrm{TN}_e)$ is introduced to measure how many independent observations $\boldsymbol{\mathcal{M}}$ is equivalent to. Furthermore, we have $r|\mathcal{M}_e \sim Be(\mathrm{TP}_e+\lambda, \mathrm{FN}_e+\lambda)$, and the variance of $\mathrm{R}_{3\times 2}$ can be rewritten as

$$\mathrm{Var}[\mathrm{R}_{3\times 2}] = \frac{r(1-r)}{\mathrm{TP}_e+\mathrm{FN}_e}. \quad (15)$$

Comparing Eqs. (14) and (15), we obtain

$$\mathrm{TP}_e+\mathrm{FN}_e = \frac{3n_+}{1+\rho_1+4\rho_2} = \frac{\sum_{j=1}^{3}\sum_{k=1}^{2}\left(\mathrm{TP}_k^{(j)}+\mathrm{FN}_k^{(j)}\right)}{1+\rho_1+4\rho_2} \quad (16)$$

According to Eq. (5), we obtain

$$mode[r|\boldsymbol{\mathcal{M}}] = \frac{\mathrm{TP}_e}{\mathrm{TP}_e+\mathrm{FN}_e} = \mathrm{R}_{3\times 2}. \quad (17)$$

On the basis of Eqs. (9), (16), and (17), TP_e and FN_e are expressed as

$$\mathrm{TP}_e = \frac{1}{1+\rho_1+4\rho_2}\sum_{j=1}^{3}\sum_{k=1}^{2}\mathrm{TP}_k^{(j)}, \quad (18)$$

$$\mathrm{FN}_e = \frac{1}{1+\rho_1+4\rho_2}\sum_{j=1}^{3}\sum_{k=1}^{2}\mathrm{FN}_k^{(j)}. \quad (19)$$

According to Eq. (6), we obtain

$$mode[p|\boldsymbol{\mathcal{M}}] = \frac{\mathrm{TP}_e}{\mathrm{TP}_e+\mathrm{FP}_e} = \mathrm{P}_{3\times 2}. \quad (20)$$

On the basis of Eqs. (8), (18) and (20), FP_e is

$$\mathrm{FP}_e = \frac{1}{1+\rho_1+4\rho_2}\sum_{j=1}^{3}\sum_{k=1}^{2}\mathrm{FP}_k^{(j)}. \quad (21)$$

Obviously, TP_e, FP_e, and FN_e contain unknown ρ_1 and ρ_2, and their relationships are

- When $\rho_1 = \rho_2 = 0$, $\mathrm{TP}_e = \sum_{j=1}^{3}\sum_{k=1}^{2}\mathrm{TP}_k^{(j)}$. FN_e and FP_e have similar forms. These forms indicate that the posterior distribution of $r|\mathcal{M}$ is equivalent to that on six independent text corpora.

- When $\rho_1 = \rho_2 = 1$, TP_e, FP_e, and FN_e are equal to the average values of all TPs, FPs, and FNs in $\mathcal{M}$, respectively. In reality, this situation indicates that the posterior distributions based on 3×2 BCV are similar to the posterior distributions on an HO validation. Repetitions have no evident contribution to the posteriors.

In fact, R could be considered as a variant of the generalization error that takes the expectation of zero-one loss on merely positive observations. Correlations ρ_1 and ρ_2 in 3×2 BCV estimator of the generalization error have been investigated (Wang et al., 2014, 2017a). The works empirically indicate $0 \leq \rho_1 \leq 1/2$ and $1/4 \leq \rho_2 \leq 1/2$, which are also applicable for the correlations in $\mathrm{R}_{3\times 2}$. To eliminate unknown ρ_1 and ρ_2 in TP_e, FN_e, and FP_e, we take their averages over the range of $0 \leq \rho_1 \leq 1/2$ and $1/4 \leq \rho_2 \leq 1/2$ regardless of the model used. Hence,

$$\begin{aligned}\mathrm{TP}_e &\approx 8\int_{0.25}^{0.5}\int_{0}^{0.5}\frac{\sum_{j=1}^{3}\sum_{k=1}^{2}\mathrm{TP}_k^{(j)}}{1+\rho_1+4\rho_2}d\rho_1 d\rho_2\\ &\approx 0.3688\sum_{j=1}^{3}\sum_{k=1}^{2}\mathrm{TP}_k^{(j)}. \end{aligned} \quad (22)$$

Similarly, we obtain

$$\mathrm{FN}_e \approx 0.3688\sum_{j=1}^{3}\sum_{k=1}^{2}\mathrm{FN}_k^{(j)}, \quad (23)$$

$$\mathrm{FP}_e \approx 0.3688\sum_{j=1}^{3}\sum_{k=1}^{2}\mathrm{FP}_k^{(j)}. \quad (24)$$

In sum, 3×2 BCV posterior distributions of P, R and F_1 are

$$P(p = t|\mathcal{M}) = \frac{t^{\mathrm{TP}_e+\lambda}(1-t)^{\mathrm{FP}_e+\lambda}}{B(\mathrm{TP}_e+\lambda, \mathrm{FP}_e+\lambda)}, \quad (25)$$

$$P(r = t|\mathcal{M}) = \frac{t^{\mathrm{TP}_e+\lambda}(1-t)^{\mathrm{FN}_e+\lambda}}{B(\mathrm{TP}_e+\lambda, \mathrm{FN}_e+\lambda)}, \quad (26)$$

$$P(f_1 = t|\mathcal{M}) = \frac{2^{\bar{a}}(1-t)^{\bar{a}-1}(2-t)^{-\bar{a}-\bar{b}}t^{\bar{b}-1}}{B(\bar{a},\bar{b})}, \quad (27)$$

where $B(\cdot,\cdot)$ is a beta function with parameters of $\bar{a} = \mathrm{FP}_e + \mathrm{FN}_e + 2\lambda$ and $\bar{b} = \mathrm{TP}_e + \lambda$. In this study, $\lambda = 1$ is used.

2.4 CIs of P, R, and F_1 Based on 3×2 BCV

On the basis of the 3×2 BCV posterior distributions of P, R, and F_1, their corresponding CIs could be derived. The CI of P with a probability $1-\alpha$ is

$$\begin{aligned}\mathrm{CI}_p = [& Be_{\frac{\alpha}{2}}(\mathrm{TP}_e+\lambda, \mathrm{FP}_e+\lambda),\\ & Be_{1-\frac{\alpha}{2}}(\mathrm{TP}_e+\lambda, \mathrm{FP}_e+\lambda)]. \end{aligned} \quad (28)$$

The CI of R is

$$\begin{aligned}\mathrm{CI}_r = [& Be_{\frac{\alpha}{2}}(\mathrm{TP}_e+\lambda, \mathrm{FN}_e+\lambda),\\ & Be_{1-\frac{\alpha}{2}}(\mathrm{TP}_e+\lambda, \mathrm{FN}_e+\lambda)]. \end{aligned} \quad (29)$$

The CI of F_1 is

$$\mathrm{CI}_{f_1} = \left[\frac{2}{2+Be'_{1-\frac{\alpha}{2}}}, \frac{2}{2+Be'_{\frac{\alpha}{2}}}\right], \quad (30)$$

where Be'_α is the α quantile of a beta-prime distribution with parameters of $\mathrm{FP}_e + \mathrm{FN}_e + 2\lambda$ and $\mathrm{TP}_e + \lambda$.

The above CIs are more accurate than the previously proposed CIs (Wang et al., 2015; Wang and Li, 2016) because the parameters in the posterior distributions are corrected via the correlations in the 3×2 BCV estimator. Take F_1 as an example. A different CI of F_1 based on 3×2 BCV is given in (Wang et al., 2015), which employs the averaged values of FPs, FNs, and TPs in $\mathcal{M}$. Their CI is a special case of Eq. (30) with $\rho_1 = \rho_2 = 1$. Their CI is more conservative, that is, the actual degree of credibility (DOC) is larger than the nominal probability $(1-\alpha)$. Nevertheless, our CI is more accurate because it could relieve the conservativity, which is shown in the following example.

Example: Consider a similar simulation in (Wang et al., 2015), which uses a classification data set with two classes. A sample is $Z = (X, Y)$ where $P(Y = 1) = P(Y = 0) = \frac{1}{2}$, and $X|Y = 0 \sim N(\mu_0, \Sigma_0)$, $X|Y = 1 \sim N(\mu_1, \Sigma_1)$. Take $\mu_0 = (0, 0)$, $\mu_1 = (0.5, 0.5)$, and $\Sigma_0 = \Sigma_1 = I_2$. The data set size is $n = 600$ and $\alpha = 0.05$. With a logistic regression algorithm, the DOC and interval length (IL) of their CI are 99.6% and 0.117. However, the DOC and IL of our CI are 94.5% and 0.0854. Obviously, our CI has a DOC closer to $1-\alpha$ and a shorter IL, indicating that our CI is more accurate.

3 Bayes Test for Comparison of Two NLP Models

For an NLP task, assume $\mathcal{A}$ is a state-of-the-art model using D_n. When a model $\mathcal{B}$ is crafted out, it is indispensable to compare it with $\mathcal{A}$ to document whether $\mathcal{B}$ performs *significantly* better than $\mathcal{A}$ by employing the following hypotheses:

$$H_0 : \nu_{\mathcal{B}} - \nu_{\mathcal{A}} \leq 0 \;\; v.s. \;\; H_1 : \nu_{\mathcal{B}} - \nu_{\mathcal{A}} > 0, \quad (31)$$

where $\nu_{\mathcal{A}}$ and $\nu_{\mathcal{B}}$ are the evaluation metrics of models $\mathcal{A}$ and $\mathcal{B}$. In this study, P, R and F_1 are considered.

We address Problem (31) with a Bayes test (Casella and Berger, 2002), which is different to previous NHST studies (Dietterich, 1998; Alpaydin, 1999; Yildiz, 2013). A Bayes test could avoid many shortcomings of NHST reasoning, such as the egregious logic error in p-value. Moreover, a Bayes test could directly compute the probabilities of the hypotheses, which help users to make a more reasonable decision. Thus, a Bayes test is increasingly preferred and recommended recently as an advanced tool to analyze the experimental results (Benavoli et al., 2016).

In this study, we propose a Bayes test that uses the 3×2 BCV posterior distributions of P, R, and F_1 to calculate the probabilities of hypotheses, denoted as $P(H_0)$ and $P(H_1)$. Then, the test infers a decision with the heuristic rules: *Accept* H_0 *iff* $P(H_0) \geq P(H_1)$*; otherwise accept* H_1.

Before elaborating the Bayes test, several necessary denotations are introduced: the $\boldsymbol{\mathcal{M}}$ of model $\mathcal{A}$ is $\boldsymbol{\mathcal{M}}_{\mathcal{A}}$; the TP_e, FN_e, and FP_e of model $\mathcal{A}$ are $\mathrm{TP}_{e,\mathcal{A}}$, $\mathrm{FN}_{e,\mathcal{A}}$, and $\mathrm{FP}_{e,\mathcal{A}}$, respectively. The p, r, and f_1 of $\mathcal{A}$ are $p_{\mathcal{A}}$, $r_{\mathcal{A}}$, and $f_{1,\mathcal{A}}$, respectively. The denotations of $\mathcal{B}$ are defined in a similar manner. Let ν denote a user-defined metric in $\{\mathrm{P}, \mathrm{R}, \mathrm{F}_1\}$. For example, if user assign R to ν, then $r_{\mathcal{A}}$ and $r_{\mathcal{B}}$ are compared.

The key point to perform a Bayes test on Problem (31) is to tackle the distribution of the difference of $\nu_{\mathcal{A}} - \nu_{\mathcal{B}}$. However, no explicit form of the distribution exists. Thus, we estimate it using the Monte-Carlo simulation. Take R as an example. Conditioned on $\boldsymbol{\mathcal{M}}_{\mathcal{A}}$ and $\boldsymbol{\mathcal{M}}_{\mathcal{B}}$, assuming $r_{\mathcal{A}}$ is independent of $r_{\mathcal{B}}$, we wish to evaluate the probability $p(r_{\mathcal{A}} - r_{\mathcal{B}} \leq 0 | \boldsymbol{\mathcal{M}}_{\mathcal{A}}, \boldsymbol{\mathcal{M}}_{\mathcal{B}})$, that is,

$$\int_0^1 \int_0^1 \mathbb{I}(r_{\mathcal{A}} - r_{\mathcal{B}} \leq 0) P(r_{\mathcal{A}} | \boldsymbol{\mathcal{M}}_{\mathcal{A}}) \cdot P(r_{\mathcal{B}} | \boldsymbol{\mathcal{M}}_{\mathcal{B}}) dr_{\mathcal{A}} dr_{\mathcal{B}}, \quad (32)$$

where $\mathbb{I}(\cdot)$ is the indicator function that has value one *iff* the enclosed condition is true and zero otherwise. Considering that no close form of Eq. (32) exists, we have to evaluate it using Monte-Carlo simulation: (a) Sample a large number of observations from $P(r_{\mathcal{A}} | \boldsymbol{\mathcal{M}}_{\mathcal{A}})$ and $P(r_{\mathcal{B}} | \boldsymbol{\mathcal{M}}_{\mathcal{B}})$, and denote them as $\{s_{i,\mathcal{A}}\}_{i=1}^{L}$ and $\{s_{i,\mathcal{B}}\}_{i=1}^{L}$; (b) approximate Eq. (32) with the empirical proportion:

$$\frac{1}{L} \sum_{i=1}^{L} \mathbb{I}(s_{i,\mathcal{A}} - s_{i,\mathcal{B}} \leq 0), \quad (33)$$

where $L = 1{,}000{,}000$ is used.

Input: Text corpus, D_n; NLP models, $\mathcal{A}$ and $\mathcal{B}$; Evaluation metric, ν;
Output: Probabilities of the hypotheses and a decision between "Accept H_0" and "Accept H_1";
Construct $\mathbb{P}$ on D_n according to Table 1;
Train and validate models $\mathcal{A}$ and $\mathcal{B}$ on $\mathbb{P}$, and summarize the results as $\boldsymbol{\mathcal{M}}_{\mathcal{A}}$ and $\boldsymbol{\mathcal{M}}_{\mathcal{B}}$, respectively;
Apply Eqs. (22), (23) and (24) on $\boldsymbol{\mathcal{M}}_{\mathcal{A}}$ and $\boldsymbol{\mathcal{M}}_{\mathcal{B}}$ to get $(\mathrm{TP}_{e,\mathcal{A}}, \mathrm{FN}_{e,\mathcal{A}}, \mathrm{FP}_{e,\mathcal{A}})$ and $(\mathrm{TP}_{e,\mathcal{B}}, \mathrm{FN}_{e,\mathcal{B}}, \mathrm{FP}_{e,\mathcal{B}})$;
if ν *is* P **then**
    $P(\nu_{\mathcal{A}} | \boldsymbol{\mathcal{M}}_{\mathcal{A}}) \leftarrow$ use Eq. (25) on $\mathrm{TP}_{e,\mathcal{A}}$ and $\mathrm{FP}_{e,\mathcal{A}}$;
    $P(\nu_{\mathcal{B}} | \boldsymbol{\mathcal{M}}_{\mathcal{B}}) \leftarrow$ use Eq. (25) on $\mathrm{TP}_{e,\mathcal{B}}$ and $\mathrm{FP}_{e,\mathcal{B}}$;
end
else if ν *is* R **then**
    $P(\nu_{\mathcal{A}} | \boldsymbol{\mathcal{M}}_{\mathcal{A}}) \leftarrow$ use Eq. (26) on $\mathrm{TP}_{e,\mathcal{A}}$ and $\mathrm{FN}_{e,\mathcal{A}}$;
    $P(\nu_{\mathcal{B}} | \boldsymbol{\mathcal{M}}_{\mathcal{B}}) \leftarrow$ use Eq. (26) on $\mathrm{TP}_{e,\mathcal{B}}$ and $\mathrm{FN}_{e,\mathcal{B}}$;
end
else if ν *is* F_1 **then**
    $P(\nu_{\mathcal{A}} | \boldsymbol{\mathcal{M}}_{\mathcal{A}}) \leftarrow$ use Eq. (27) on $\mathrm{TP}_{e,\mathcal{A}}$, $\mathrm{FP}_{e,\mathcal{A}}$ and $\mathrm{FN}_{e,\mathcal{A}}$;
    $P(\nu_{\mathcal{B}} | \boldsymbol{\mathcal{M}}_{\mathcal{B}}) \leftarrow$ use Eq. (27) on $\mathrm{TP}_{e,\mathcal{B}}$, $\mathrm{FP}_{e,\mathcal{B}}$ and $\mathrm{FN}_{e,\mathcal{B}}$;
end
Approximate $P(\nu_{\mathcal{A}} - \nu_{\mathcal{B}} \leq 0 | \boldsymbol{\mathcal{M}}_{\mathcal{A}}, \boldsymbol{\mathcal{M}}_{\mathcal{B}})$ with Monte-Carlo simulation (refer to Eq. (33));
$P(H_0) \leftarrow P(\nu_{\mathcal{A}} - \nu_{\mathcal{B}} \leq 0 | \boldsymbol{\mathcal{M}}_{\mathcal{A}}, \boldsymbol{\mathcal{M}}_{\mathcal{B}})$;
$P(H_1) \leftarrow 1 - P(\nu_{\mathcal{A}} - \nu_{\mathcal{B}} \leq 0 | \boldsymbol{\mathcal{M}}_{\mathcal{A}}, \boldsymbol{\mathcal{M}}_{\mathcal{B}})$;
if $P(H_0) \geq P(H_1)$ **then**
    Return $(P(H_0), P(H_1)$, "Accept H_0");
end
else
    Return $(P(H_0), P(H_1)$, "Accept H_1");
end

Algorithm 1: A Bayes test for comparing P, R and F_1 of two NLP models.

On the basis of the above analysis, the sketch of Bayes test is shown in Algorithm 1. The algorithm performs hypothesis testing procedures for P, R, and F_1 according to the specific value of ν. When different evaluation metrics are used, the corresponding hypothesis testing problems (refer to Problem (31)) are different, and the decisions might be different but reasonable, even though the same text corpus is used in these problems. Thus, the Bayes test helps users to investigate the difference of $\mathcal{A}$ and $\mathcal{B}$ with different perspectives and in a fine-grained manner.

Bayes test and NHST are two different types of significant tests from two philosophies: Bayesian and frequentist inferences. When the distribution of an evaluation metric is available, the Bayes test may provide more informative inferences and conclusions than the NHST. Until now, no mature and fair criterion to compare Bayes test and NHST exists. Therefore, in this study, an objective comparison between them is not provided. Instead, we show three experiments to illustrate the validity of the Bayes test.

4 Experiments and Analysis

The experiments concentrate on chunking tasks [1]. Chunking is an important task in NLP, which includes Chinese word segmentation (CWS) and NER. A chunking task could be formulated into a sequence labeling problem and addressed by employing a tag set, such as IOB2 and IOBES (Kudo and Matsumoto, 2001; Shen and Sarkar, 2005), and a widely used algorithm, such as conditional random fields (CRFs) (Lafferty et al., 2001) and LSTM (Hochreiter and Schmidhuber, 1997; Lample et al., 2016).

In this section, we perform the Bayes test on NLP chunking models with different tag sets to answer a question: could a fine-grained tag set improve the performance of a chunking model?

A chunking model is usually evaluated in terms of the metrics of P, R, and F_1. When computing them, TP indicates the count of correctly predicted chunks, FN is the count of golden chunks that are incorrectly predicted, and FP is the count of predicted chunks that are not correct.

Three different chunking tasks are considered:

CWS task: Identify a reasonable word sequence in a raw sentence. A word is regarded as a chunk, and every character in the sentence enters into a chunk. Bakeoff-2005 CWS PKU training corpus is used as D_n.

NER task: Identify the boundaries of all NER chunks without recognizing their types. CoNLL 2003 English NER training set is used as D_n, which contains four types of NER, namely, "PER", "LOC", "ORG", and "MISC". Word is used as a tagging unit, and considerable out-of-chunk words exist.

ORG task: Identify only "ORG" entities. The corpus is the same to the NER task. The count of "ORG" chunks is remarkably smaller than the count of NER chunks in the NER task, and the out-of-chunk words dominate the corpus.

In the above three tasks, CRFs are used as the sequence labeling algorithm. Other algorithms will be studied in future research.

4.1 CWS Task: "BMES" Versus "BB_2B_3MES"

The CWS task is formulated into a sequence labeling problem at character level. Two different tag sets of "BMES" and "BB_2B_3MES" are considered, which correspond to models $\mathcal{A}$ and $\mathcal{B}$, respectively. "BB_2B_3MES" is a fine-grained set that introduces two additional tags of "B_2" and "B_3" on the basis of "BMES". Zhao et al. (2006) illustrated that model $\mathcal{B}$ improves $\mathcal{A}$ without investigating the significance, which is performed here.

Task	ν	Tag set 1 (%)	Tag set 2 (%)
CWS		BMES	BB_2B_3MES
	P	[95.55, 95.62]	[95.60, 95.67]
	R	[95.04,95.11]	[95.16,95.23]
	F_1	[95.30,95.36]	[95.39,95.44]
NER		IOB2	IOBES
	P	[90.59, 91.30]	[90.70,91.41]
	R	[87.69,88.48]	[87.78, 88.57]
	F_1	[89.21,89.77]	[89.32,89.87]
ORG		IOB2	IOBES
	P	[91.37,92.86]	[91.85,93.31]
	R	[64.89,67.11]	[64.45,66.68]
	F_1	[76.06,77.74]	[75.93,77.61]

Table 2: CIs of the three tasks ($\alpha = 0.05$).

ν	$P(H_0)$	$P(H_1)$	Decision
P	0.024	0.976	Accept H_1
R	0.001	0.999	Accept H_1
F_1	0.001	0.999	Accept H_1

Table 3: Decisions of the Bayes test in the CWS task.

In the task, the unigram, bigram, and trigram of characters are used as features, and their windows are [-2,2]. The 3×2 BCV posterior distributions of

[1] The code for the experiments in this paper is found on: https://github.com/RamboWANG/acl2019

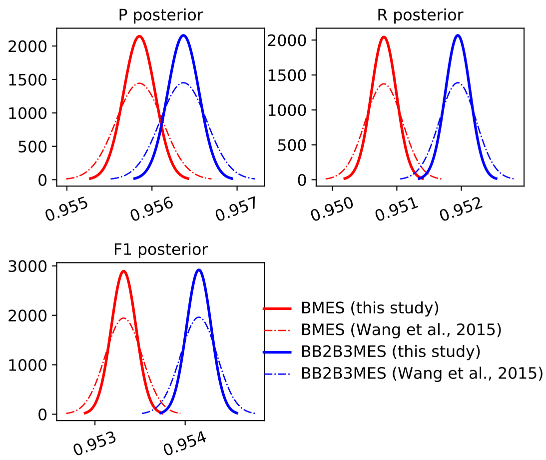

Figure 1: 3×2 BCV posterior distributions in the CWS task.

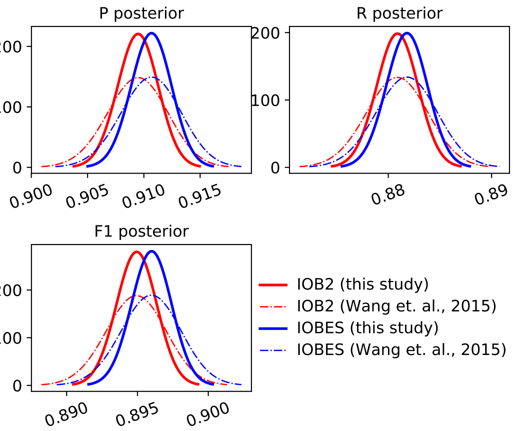

Figure 2: 3×2 BCV posterior distributions in the NER task.

the two CWS models are given in Figure 1, and the CIs in $\alpha = 0.05$ are given in Table 2. Each curve ranges from 0.001 quantile to 0.999 quantile. The curves in solid lines correspond to Eqs. (25), (26), and (27), which are recommended in this study.

Two observations are concluded from Figure 1. First, our proposed posterior distributions, which yield more accurate CIs, are taller and thinner than those in (Wang et al., 2015). Second, the posterior distributions of the R and F_1 between models $\mathcal{A}$ and $\mathcal{B}$ have smaller overlaps than those of P. The smaller overlap indicates that the additional tags of "B_2" and "B_3" mainly improve the R and F_1 of the CWS model.

The Bayes test is performed on $\mathcal{A}$ and $\mathcal{B}$. The probabilities of the hypotheses and decisions are given in Table 3. H_1 holds in the probability of 0.98 for P, whereas H_1 holds in the probabilities of approximately 1 for R and F_1. Table 3 illustrates that the fine-grained tag set significantly improves the CWS model, and the improvements in R and F_1 are larger than P.

4.2 NER Task: "IOB2" Versus "IOBES"

In this task, word and POS are used as features. The unigram, bigram, and trigram of word and POS are included in the feature template. The window size of each type of feature is set to [-2,2]. "IOBES" in model $\mathcal{B}$ is a fine-grained tag set, which adds tags "E" and "S" to "IOB2" in $\mathcal{A}$.

Posterior distributions of P, R, and F_1 of models $\mathcal{A}$ and $\mathcal{B}$ are given in Figure 2. The posterior distributions of the two models have large overlaps, which indicate that the improvement in $\mathcal{B}$ is not evident. Corresponding CIs are given in Table 2. The CIs of the two models have also large overlaps, indicating the insignificant differences between the two models. Table 4 presents the decisions of the Bayes test, which are identical on the three metrics, that is, "Accept H_1". However, the improvement is not remarkable because $P(H_1)$ are lower than 0.8. Moreover, the fine-grained tag set, "IOBES," exerts more effort to improve P than R because $P(H_1) = 0.68$ for P is larger than $P(H_1) = 0.63$ for R.

ν	$P(H_0)$	$P(H_1)$	Decision
P	0.321	0.679	Accept H_1
R	0.372	0.628	Accept H_1
F_1	0.300	0.700	Accept H_1

Table 4: Decisions of the Bayes test in the NER task.

4.3 ORG Task: "IOB2" Versus "IOBES"

In this task, the settings of features are the same with the NER task. However, the distributions of tags become more skewed than those of the NER task, that is, tag "O" possesses a larger proportion. Thus, the decisions of the Bayes test are remarkably different. Specifically, the posterior distributions of P, R, and F_1 are given in Figure 3, which indicate that the improvement in $\mathcal{B}$ is not evident. Surprisingly, for R and F_1, the posterior distribution of $\mathcal{B}$ shifts to the left of that of $\mathcal{A}$, which illustrates the fine-grained tag set, namely, "IOBES," deteriorates R and F_1. A possible reason is the fine-grained tag set, namely, "IOBES," leads to

ν	$P(H_0)$	$P(H_1)$	Decision
P	0.191	0.809	Accept H_1
R	0.706	0.294	Accept H_0
F_1	0.587	0.413	Accept H_0

Table 5: Decisions of the Bayes test in the ORG task.

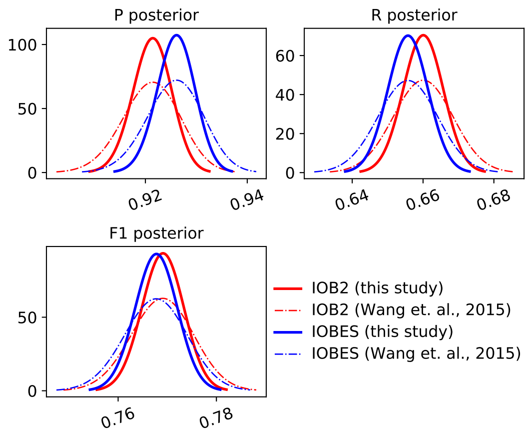

Figure 3: 3×2 BCV posterior distributions in the ORG task.

more skewed proportions of tags than "IOB2."

The decisions of the Bayes test are given in Table 5. The probability of the improvement to P exceeds 0.8, that is, $P(H_1) = 0.81$. However, the fine-grained tag set harms R and F_1 in a sense of $P(H_0) = 0.71$ for R and $P(H_0) = 0.59$ for F_1.

The above three tasks illustrate the validity of the Bayes test, which provide accurate CIs of P, R, and F_1 and the estimation of $P(H_0)$ and $P(H_1)$. The results are more informative for interpretations and help to make a reliable decision.

5 Related Work

Over the last few past decades, many studies have contributed to validate whether the standard significant tests are adequate for comparing NLP models (Gillick and Cox, 1989; Yeh, 2000; Daelemans and Hoste, 2002; Koehn, 2004; Riezler and Maxwell, 2005; Berg-Kirkpatrick et al., 2012; Søgaard, 2013; Søgaard et al., 2014; Névéol et al., 2016; Dror et al., 2017, 2018). These studies observed that standard tests tend to infer invalid comparison conclusions. Two important questions arise from the observations: 1) How to correctly perform CV for NLP model comparison? 2) What are the distributions of the common evaluation metrics in NLP, such as P, R, and F_1?

The first question could refer to many studies in machine learning, which investigated various CV methods in algorithm comparison, including repeated learning-testing (Nadeau and Bengio, 2003; Wang et al., 2019), K-fold CV (Kohavi et al., 1995; Rodríguez et al., 2010, 2013; Moreno-Torres et al., 2012), 5×2 CV (Dietterich, 1998; Alpaydin, 1999; Yildiz, 2013), and $m \times 2$ BCV (Wang et al., 2014, 2015, 2017a,b). In these studies, $m \times 2$ BCV might be a better option for comparing NLP models because it leads to stable estimation of evaluation metrics and the $m \times 2$ BCV tests possesses higher powers and replicabilities (Wang et al., 2014, 2017b). Moreover, on a text corpus, certain frequency distributions over linguistic units between training and validation sets in two-fold CV intuitively possess smaller divergence than those in five-fold or ten-fold CV. Therefore, $m \times 2$ BCV should be investigated when comparing NLP models.

The second question is pioneered in the work of (Goutte and Gaussier, 2005), which proved the posterior distributions of P and R in an HO validation. The posterior distributions make an exact comparison possible (Zhang and Su, 2012; Wang and Li, 2016). However, the distribution of F_1 is difficult to tackle, because it is a complex function. Zhang et al. (2015a,b, 2016) employed complicated probabilistic graphic representations and Bayesian hierarchical models to estimate and compare F_1 measures. Fortunately, Wang et al. (2015) obtained an exact close-form of posterior distribution of F_1, which is a function with regard to a beta-prime distribution. These studies provided a rigorous theoretical guarantee for pursuing the 3×2 BCV posterior distributions of P, R, and F_1.

6 Conclusions and Future Work

In this study, we obtained accurate posterior distributions of P, R, and F_1 on the basis of a 3×2 BCV, which is an essential part in conducting the comparison of two NLP models. On the basis of the posterior distributions, a Bayes test is proposed, which provides the probabilities of the hypotheses and help users to make a reasonable decision. Finally, three experiments on chunking tasks are performed to illustrate the validity of the Bayes test. For NLP practitioners, we recommend here three guidelines:

(1) A t-test should be avoided in a comparison of two NLP models on the basis of the precision, recall and F_1 measure.

(2) The 3×2 BCV could be preferred to evaluate the performance of an NLP model in the task of model comparison.

(3) The Bayes test on the basis of the 3×2 BCV could provide informative and fine-grained measures of the differences of precisions, recalls and F_1 measures of two NLP models, and the measures could help practitioners to make a reasonable decision.

In the future, we will refine the Bayes test of P, R, and F_1 in an $m \times 2$ BCV and provide accurate interval estimation of other evaluation metrics on the basis of the confusion matrix. Obtaining the posterior distribution of an evaluation metric of a model is still a key problem in this valuable research area.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work is supported by the National Social Science Foundation of China under Grants no. 16BTJ034. The experiments are supported by High Performance Computing System of Shanxi University. The corresponding author is Jihong Li.

References

Ethem Alpaydin. 1999. Combined 5×2 cv f-test for comparing supervised classification learning algorithms. *Neural Computation*, 11(8):1885–1892.

Alessio Benavoli, Giorgio Corani, Janez Demsar, and Marco Zaffalon. 2016. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005. Association for Computational Linguistics.

George Casella and Roger L Berger. 2002. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Walter Daelemans and Véronique Hoste. 2002. Evaluation of machine learning methods for natural language processing tasks. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 755–760.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhikers guide to testing statistical significance in natural language processing. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1383–1392. Association for Computational Linguistics.

Laurence Gillick and Stephen J Cox. 1989. Some statistical issues in the comparison of speech recognition algorithms. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 532–535. IEEE.

Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European Conference on Information Retrieval*, pages 345–359. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.

Ron Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machines. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

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 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Jihong Li, Ruibo Wang, Weilin Wang, Bo Gu, and Guochen Li. 2009. Automatic labeling of semantic role on chinese framenet using conditional random fields. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03*, pages 259–262. IEEE Computer Society.

Jose G Moreno-Torres, Jos A Sez, and Francisco Herrera. 2012. Study on the impact of partition-induced dataset shift on k-fold cross-validation. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(8):1304–1312.

Claude Nadeau and Yoshua Bengio. 2003. Inference for the generalization error. *Machine Learning*, 52(3):239–281.

Aurélie Névéol, Kevin Cohen, Cyril Grouin, and Aude Robert. 2016. Replicability of research in biomedical natural language processing: a pilot evaluation for a coding task. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 78–84.

Stefan Riezler and John T Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for mt. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 57–64.

Juan D Rodríguez, Aritz Pérez, and Jose A Lozano. 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):569–575.

Juan D Rodríguez, Aritz Pérez, and Jose A Lozano. 2013. A general framework for the statistical analysis of the sources of variance for classification error estimators. *Pattern Recognition*, 46(3):855–864.

Hong Shen and Anoop Sarkar. 2005. Voting between multiple data representations for text chunking. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 389–400. Springer.

Anders Søgaard. 2013. Estimating effect size across datasets. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 607–611.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Héctor Martínez Alonso. 2014. What's in a p-value in nlp? In *Proceedings of the eighteenth conference on computational natural language learning*, pages 1–10.

Ruibo Wang, Jihong Li, Xingli Yang, and Jing Yang. 2019. Block-regularized repeated learning-testing for estimating generalization error. *Information Sciences*, 477:246–264.

Ruibo Wang, Yu Wang, Jihong Li, Xingli Yang, and Jing Yang. 2017a. Block-regularized $m \times 2$ cross-validated estimator of the generalization error. *Neural Computation*, 29(2):519–554.

Yu Wang and Jihong Li. 2016. Credible intervals for precision and recall based on a k-fold cross-validated beta distribution. *Neural Computation*, 28(8):1694–1722.

Yu Wang, Jihong Li, and Yanfang Li. 2017b. Choosing between two classification learning algorithms based on calibrated balanced 5×2 cross-validated f-test. *Neural Processing Letters*, 46(1):1–13.

Yu Wang, Jihong Li, Yanfang Li, Ruibo Wang, and Xingli Yang. 2015. Confidence interval for f_1 measure of algorithm performance based on blocked 3x2 cross-validation. *IEEE Transactions on Knowledge & Data Engineering*, 27(3):651–659.

Yu Wang, Ruibo Wang, Huichen Jia, and Jihong Li. 2014. Blocked 3×2 cross-validated t-test for comparing supervised classification learning algorithms. *Neural Computation*, 26(1):208–235.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 947–953. Association for Computational Linguistics.

Olcay Taner Yildiz. 2013. Omnivariate rule induction using a novel pairwise statistical test. *IEEE Transactions on Knowledge and Data Engineering*, 25(9):2105–2118.

Dell Zhang, Jun Wang, Emine Yilmaz, Xiaoling Wang, and Yuxin Zhou. 2016. Bayesian performance comparison of text classifiers. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 15–24. ACM.

Dell Zhang, Jun Wang, and Xiaoxue Zhao. 2015a. Estimating the uncertainty of average f_1 scores. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, pages 317–320. ACM.

Dell Zhang, Jun Wang, Xiaoxue Zhao, and Xiaoling Wang. 2015b. A bayesian hierarchical model for comparing average f_1 scores. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 589–598. IEEE.

Peng Zhang and Wanhua Su. 2012. Statistical inference on recall, precision and average precision under random selection. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, pages 1348–1352. IEEE.

Hai Zhao, Chang-Ning Huang, and Mu Li. 2006. An improved chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 162–165.

A Proof of Eq. (4)

According to Eqs. (1) and (3), we obtain

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \frac{2\mathrm{TP}}{n_+}. \quad (34)$$

Because $\mathrm{TP}|\mathrm{TP}+\mathrm{FN}$ follows a binomial distribution with parameters of $n_+/2$ and r and $\mathrm{TP}+\mathrm{FN} = n_+/2$ is a constant, we obtain

$$\mathrm{Var}[\mathrm{TP}] = \frac{n_+}{2} r(1-r). \quad (35)$$

Based on Eq. (34), we know

$$\begin{aligned}\mathrm{Var}[\mathrm{R}] &= \mathrm{Var}[\frac{2\mathrm{TP}}{n_+}] = \frac{4}{n_+^2}\mathrm{Var}[\mathrm{TP}] \\ &= \frac{2r(1-r)}{n_+}.\end{aligned}$$

■

B Proof of Eq. (14)

According to Eq. (11), $\mathrm{Var}[\mathrm{R}_{3\times 2}]$ can be decomposed into

$$\begin{aligned}\mathrm{Var}[\mathrm{R}_{3\times 2}] &= \mathrm{Var}\left[\frac{1}{3}\sum_{j=1}^{3}\mathrm{R}^{(j)}\right] \\ &= \frac{1}{9}\Bigg\{\sum_{j=1}^{3}\mathrm{Var}\left[\mathrm{R}^{(j)}\right] \\ &+\sum_{\substack{j=1}}^{3}\sum_{\substack{j'=1 \\ j\neq j'}}^{3}\mathrm{Cov}\left[\mathrm{R}^{(j)},\mathrm{R}^{(j')}\right]\Bigg\}. \quad (36)\end{aligned}$$

Assume $\mathrm{Var}\left[\mathrm{R}^{(j)}\right]$ doesn't depend on the particular realization of $\mathbf{P}_j$, then $\mathrm{Var}\left[\mathrm{R}^{(j)}\right]$ for all j are identical. Furthermore, since the number of overlapping samples between the two training sets in $\mathbf{P}_j$ and $\mathbf{P}_{j'}$ equals to $n/4$ with $j \neq j'$, we could reasonably assume $\mathrm{Cov}\left[\mathrm{R}^{(j)},\mathrm{R}^{(j')}\right]$ for all $j \neq j'$ are identical and independent to j and j'. Thus, we obtain

$$\begin{aligned}\mathrm{Var}[\mathrm{R}_{3\times 2}] &= \frac{1}{3}\Bigg\{\mathrm{Var}\left[\mathrm{R}^{(j)}\right] \\ &+2\mathrm{Cov}\left[\mathrm{R}^{(j)},\mathrm{R}^{(j')}\right]\Bigg\}. \quad (37)\end{aligned}$$

Since $\mathrm{R}^{(j)} = (\mathrm{R}_1^{(j)} + \mathrm{R}_2^{(j)})/2$, assume $\mathrm{Var}\left[\mathrm{R}_1^{(j)}\right] = \mathrm{Var}\left[\mathrm{R}_2^{(j)}\right]$, we have

$$\begin{aligned}\mathrm{Var}\left[\mathrm{R}^{(j)}\right] &= \mathrm{Var}\left[\frac{1}{2}(\mathrm{R}_1^{(j)}+\mathrm{R}_2^{(j)})\right] \\ &= \frac{1}{2}\left\{\mathrm{Var}\left[\mathrm{R}_k^{(j)}\right]+\mathrm{Cov}\left[\mathrm{R}_k^{(j)},\mathrm{R}_{k'}^{(j)}\right]\right\} \quad (38)\end{aligned}$$

where $k \neq k'$. Furthermore, according to Eq. (4) and the definition of ρ_1, we obtain

$$\mathrm{Var}\left[\mathrm{R}_k^{(j)}\right] = 2r(1-r)/n_+, \quad (39)$$

$$\mathrm{Cov}\left[\mathrm{R}_k^{(j)},\mathrm{R}_{k'}^{(j)}\right] = 2\rho_1 r(1-r)/n_+. \quad (40)$$

Substituting Eqs. (39) and (40) into Eq. (38), we obtain

$$\mathrm{Var}\left[\mathrm{R}^{(j)}\right] = \frac{1+\rho_1}{n_+}r(1-r). \quad (41)$$

Similarly, assume $\mathrm{Cov}\left[\mathrm{R}_k^{(j)},\mathrm{R}_{k'}^{(j')}\right]$ doesn't depend on k and k', then

$$\mathrm{Cov}\left[\mathrm{R}^{(j)},\mathrm{R}^{(j')}\right] = \mathrm{Cov}\left[\mathrm{R}_k^{(j)},\mathrm{R}_{k'}^{(j')}\right], \quad (42)$$

where $k, k' = 1, 2$. According to the definition of ρ_2, we obtain

$$\mathrm{Cov}\left[\mathrm{R}^{(j)},\mathrm{R}^{(j')}\right] = \frac{2\rho_2}{n_+}r(1-r). \quad (43)$$

Substituting Eqs. (41) and (43) into Eq. (37), we obtain

$$\mathrm{Var}[\mathrm{R}_{3\times 2}] = \frac{1+\rho_1+4\rho_2}{3n_+}r(1-r).$$

■