Appendix I. Learning Classifier Parameters to Optimize the F-measure

The F-measure is defined as

$$F(\beta) = \frac{(\beta^2+1)pr}{\beta^2 p + r} = \left(\frac{\beta^2}{(\beta^2+1)}\frac{1}{r} + \frac{1}{(\beta^2+1)}\frac{1}{p}\right)^{-1}$$

Therefore, to maximize the F-measure is equivalent to minimizing the risk

$$\begin{aligned}\frac{\beta^2}{(\beta^2+1)}\frac{1}{r} + \frac{1}{(\beta^2+1)}\frac{1}{p} &= \frac{\beta^2}{(\beta^2+1)}\frac{n_{11}+n_{12}}{n_{11}} + \frac{1}{(\beta^2+1)}\frac{n_{11}+n_{21}}{n_{11}} \\ &= 1 + \frac{\beta^2 n_{12} + n_{21}}{(\beta^2+1)n_{11}}\end{aligned}$$

or the risk function:

$$R \equiv \frac{\beta^2 n_{12} + n_{21}}{(\beta^2+1)n_{11}} = \frac{1}{(\beta^2+1)}\frac{\beta^2 n_{12} + n_{21}}{n_1 - n_{12}} = \frac{\alpha_{12}n_{12} + \alpha_{21}n_{21}}{n_1 - n_{12}}$$

where

$$\alpha_{21} \equiv \frac{1}{(\beta^2+1)}, \quad \alpha_{12} \equiv \frac{\beta^2}{(\beta^2+1)}, \quad (\alpha_{12} + \alpha_{21} = 1)$$

When β=1, it is equivalent to prefer a small $\frac{n_{12}+n_{21}}{n_{11}}$ ratio.

The misclassification counts n_{12} and n_{21} can be approximated with the following differentiable loss functions l_{12} and l_{21} assuming that $c(\mathbf{x})$ refers to the real class label of the input feature vector and the class label assigned by the classifier is determined by a discrimination function $g(\mathbf{x})$:

$$\begin{aligned}n_{12} &\approx l_{12} = \sum_{\mathbf{x}:c(\mathbf{x})=1, g(\mathbf{x})<0} l(-d_{\mathbf{x}}) \\ n_{21} &\approx l_{21} = \sum_{\mathbf{x}:c(\mathbf{x})=2, g(\mathbf{x})\geq 0} l(+d_{\mathbf{x}})\end{aligned}$$

where $l(d_{\mathbf{x}})$ is an approximation to a zero-one loss function, as a function of the misclassification distance $d_{\mathbf{x}}$:

$$l(d_{\mathbf{x}}) = \frac{1}{\pi}\tan^{-1}\left(\frac{d_{\mathbf{x}}}{d_0}\right) + \frac{1}{2}$$

and the distance function is defined in terms of the normalized discrimination function:

$$\begin{aligned}
d_{\mathbf{x}} &= \frac{g(\mathbf{x},\ \mathbf{w},\ \Lambda)}{\|\mathbf{w}\|} = \frac{\mathbf{w}\bullet\mathbf{s}}{\|\mathbf{w}\|} \\
\mathbf{s} &= \left[\ldots,\ \log f(x_i|\mathbf{C}) - \log f(x_i|\overline{\mathbf{C}}),\ldots,\ \log P(\mathbf{C}) - \log P(\overline{\mathbf{C}})\right] \\
\mathbf{w} &= \left[w_1, w_2, \ldots, w_D, w_0\right]
\end{aligned}$$

where the weights are introduced for scaling the log-likelihood ratios which have large dynamic ranges. Initially, all weighting factors are set to 1.

In other words, we have:

$$\begin{aligned}
\nabla R &= \nabla\frac{\alpha_{12}l_{12} + \alpha_{21}l_{21}}{n_1 - l_{12}} \\
&\approx \frac{(n_1 - n_{12})(\alpha_{12}\nabla l_{12} + \alpha_{21}\nabla l_{21}) + (\alpha_{12}n_{12} + \alpha_{21}n_{21})\nabla l_{12}}{(n_1 - n_{12})^2} \\
&= \frac{(\alpha_{12}n_1 + \alpha_{21}n_{21})\nabla l_{12} + \alpha_{21}(n_1 - n_{12})\nabla l_{21}}{(n_1 - n_{12})^2} \equiv k_{12}\nabla l_{12} + k_{21}\nabla l_{21}
\end{aligned}$$

where

$$\begin{aligned}
k_{21} &\equiv \frac{\alpha_{21}}{(n_1 - n_{12})} = \frac{1}{\beta^2 + 1}\frac{1}{(n_1 - n_{12})} \\
k_{12} &\equiv \frac{(\alpha_{12}n_1 + \alpha_{21}n_{21})}{(n_1 - n_{12})^2} = \frac{1}{\beta^2 + 1}\frac{(\beta^2 n_1 + n_{21})}{(n_1 - n_{12})^2}
\end{aligned}$$

Since

$$\begin{aligned}
\nabla l_{12} &= \sum_{\mathbf{x}:c(\mathbf{x}) = 1, g(\mathbf{x}) < 0}\nabla l(-d_{\mathbf{x}}) = -\sum l'(-d_{\mathbf{x}})\nabla d_{\mathbf{x}} \\
\nabla l_{21} &= \sum_{\mathbf{x}:c(\mathbf{x}) = 2, g(\mathbf{x}) \geq 0}\nabla l(+d_{\mathbf{x}}) = +\sum l'(+d_{\mathbf{x}})\nabla d_{\mathbf{x}}
\end{aligned}$$

and

$$l'\left(+d_{\mathbf{x}}\right) = \frac{1}{\pi}\frac{d_0}{d_{\mathbf{x}}^2 + d_0^2} = l'\left(-d_{\mathbf{x}}\right)$$

the gradient of the risk function is equivalent to

$$\begin{aligned} \nabla R &\approx k_{12}\nabla l_{12} + k_{21}\nabla l_{21} \\ &= k_{21}\left[\sum_{\substack{c(\vec{\mathbf{x}})=2 \\ g(\mathbf{x}) \geq 0}} l'\left(+d_{\mathbf{x}}\right)\nabla d_{\mathbf{x}}\right] - k_{12}\left[\sum_{\substack{c(\mathbf{x})=1 \\ g(\vec{\mathbf{x}}) < 0}} l'\left(-d_{\mathbf{x}}\right)\nabla d_{\mathbf{x}}\right] \end{aligned}$$

The evaluation of $\nabla d_{\mathbf{x}}$ is the same as for the WPR optimization criterion.

Appendix II. Statistical Significance Test on Performance Improvement

To see whether the relatively small improvement for the bigrams in the iteratively integrated system of the Chinese unknown word extraction task is statistically significant [Wonnacott 90] compared with the trigram and quadragram cases, we divide the 1,000 sample sentences into 10 groups (with 100 sentences per group), and estimate their respective performance (i.e., precision, recall, WPR and FM). The differences in performance, $D_i \equiv X_i - Y_i$ (i=1, 10), between the first and the final iterations are estimated (X_i: performance for the *i*-th group of sentences after iteratively refining the augmented dictionary, Y_i: performance before applying iterative refinement). The average difference ($\overline{D} \equiv \overline{X} - \overline{Y}$) is then compared with the sample variance (s_D^2) of the differences to see whether such differences are statistically significant. More precisely, we made a hypothesis test to reject the hypothesis that there is no difference in performance (i.e., $\mu_D = 0$) between the first and the final iterations based on the following testing statistic:

$$T = \frac{\overline{D} - \mu_D}{\sqrt{\sigma_D^2 / K}}$$

$$\overline{D} = \frac{1}{K} \sum_{i=1}^{K} D_i \qquad (K = 10)$$

$$D_i = X_i - Y_i \qquad (i = 1, K)$$

$$\sigma_D^2 \approx s_D^2 = \frac{1}{K-1} \sum_{i=1}^{K} \left(D_i - \overline{D} \right)^2$$

where D is the random variable denoting the difference in performance, $\overline{D}$ is its sample mean of all the estimated differences D_i (for the *i*-th group of sentences), μ_D is the population mean of the difference, σ_D^2 is the population variance, and K is the number of groups for estimating the differences (K=10 in the current test). In the above testing, σ_D^2 is approximated by its sample variance s_D^2, and D_i are assumed to be independently and identically distributed with mean μ_D

and variance σ_D^2. With K=10, T can be approximated as a student-t distribution with 9 degrees of freedom. To reject the hypothesis with more than 95% confidence, the T value must be greater than 1.83 when the hypothesis holds (i.e., when $\mu_D = 0$) using the observed data. In fact, we have T= 4.248, 9.142, 7.439, and 7.436 for bigram precision, recall, WPR and FM, respectively, which all exceed the threshold 3.89 for 99.8% confidence. We can therefore conclude that the smaller improvement for bigrams is not due to the statistical variation in estimating the performance, but due to the effect of the proposed iterative algorithm.