Performance Metrics for Word Sense Disambiguation

Trevor Cohn
Department of Computer Science and Software Engineering
University of Melbourne, VIC 3010, Australia
email: tacohn@cs.mu.oz.au   fax: +61-3-9348-1184

Abstract

This paper presents the area under the Receiver Operating Characteristics (ROC) curve as an alternative metric for evaluating word sense disambiguation performance. The current metrics – accuracy, precision and recall – while suitable for two-way classification, are shown to be inadequate when disambiguating between three or more senses. Specifically, these measures do not facilitate comparison with baseline performance nor are they sensitive to non-uniform misclassification costs. Both of these issues can be addressed using ROC analysis.

1 Introduction

Word sense disambiguation (WSD) is one of the large open problems in the field of natural language processing, and in recent years has attracted considerable research interest (Ide and Veronis, 1998). The increasing availability of large corpora along with electronic sense inventories (such as WordNet; Fellbaum (1998)) has permitted the application of a raft of machine learning techniques to the task and provided an empirical means of performance evaluation. Until recently, most performance evaluation was conducted on disparate data sets, with only the *line* and *interest* corpora being used in a significant number of studies (Leacock et al., 1993; Bruce and Wiebe, 1994). SENSEVAL, a global evaluation performed in 1998 (Kilgarriff, 1998) and again in 2001 (Edmonds and Cotton, 2001), provided a common set of disambiguation tasks and performance evaluation criteria, allowing an objective comparison between competing methods.

These workshops included the tasks of disambiguating all words in a given text (the all-words task), and disambiguating each occurrence of a given word when it appears with a short context of a few surrounding sentences (the lexical sample task). Performance in the two tasks was measured in terms of precision and recall. Precision was defined as the proportion of classified instances that were correctly classified, and recall as the proportion of instances classified correctly – these allow for the possibility of an algorithm choosing not to classify a given instance. This evaluation criterion is insensitive to both the type of misclassification (is the predicted sense more closely related to the correct sense than other possible senses?) and the confidence with which the classifier has made the prediction (is the correct sense allocated a high probability despite not being given the highest value by the classifier?).

These problems led Resnik and Yarowsky (1999) to suggest an evaluation metric to provide partial credit for incorrectly classified instances. They penalise probability mass assigned to incorrect senses weighted by what they term the communicative/semantic distance between the that predicted sense and the correct sense. Using such measures, systems that confuse homographs would be penalised most heavily, while those that confuse fine-grained senses would only attract a minor penalty. The score assigned to a particular algorithm is highly reliant on the distances between senses; altering the relative penalties may well promote a previously non-optimal classifier to be the best performing classifier.

In order to highlight the problems in the existing evaluation methods, it is worth clarifying the qualities such a method should possess. Ideally, the evaluation metric should provide the following features:

(1) allow comparison of the performance of two or more classifiers on the same problem, ranking them in order of quality of prediction.

(2) penalise incorrectly classified instances based on the distance, or confusability between the predicted and correct sense, when disambiguating between three or more sentences. These penalties are henceforth referred to as (non-uniform) misclassification costs.

(3) allow comparison to baseline performance – that of the classifier which always predicts only the *a priori* majority sense.

(4) provide a readily interpretable measure of performance.

This paper analyses the metrics that have been used in assessing WSD performance in light of the above criteria. An alternative metric, Receiver Operating Characteristics (ROC), is proposed and shown to have favourable properties with respect to the criteria. Section 2 describes the shortcomings of the current metrics. Section 3 shows how ROC analysis can be applied to WSD evaluation. Section 4 provides a discussion in the context of empirical studies and I conclude in section 5 with thoughts for future study.

2 Problem Statement

Many comparisons of WSD performance use predictive accuracy as the sole means of comparison. Accuracy is defined as the proportion of instances that were disambiguated correctly, and is often compared to a baseline – the performance of the classifier that predicts the majority sense for every instance. Baseline performance varies greatly between words: from lower than 10% to greater than 90%. Without some form of normalisation, comparison of the results of different classifiers on different problems is impossible. The kappa statistic (Carletta, 1996) may be used to normalise accuracy, adjusting the result for the expected agreement with the perfect classifier by chance, thus satisfying criterion (3).

Implicit in the use of accuracy is the assumption that misclassification costs are equal (or equivalently, the set of senses are all equally similar to one another). Dictionary definitions and indeed, linguistic intuitions, tell us that some sense pairs are more closely related than others. A number of dictionaries present sense hierarchies for words based on their similarities. The guidelines used by lexicographers to determine what constitutes a homograph or sense vary considerably between dictionaries. Even individual lexicographers differ in their systematic preferences as to whether they conflate similar senses into one ('lumpers') or present them as a disparate set ('splitters') (Kilgarriff, 1997; Landau, 2001). Depending on the dictionary's purpose, factors such as frequency of occurrence, semantic and syntactic similarity, pronunciation and etymology of a given word are considered (with differing priority) when identifying word's senses. Accordingly, sense definitions are rarely compatible between different dictionaries (or thesauri), presenting issues for WSD tasks using only a single source as the sense inventory.

For a binary disambiguation task, misclassification costs should be uniform – we would not expect the cost of misclassifying an instance of sense$_a$ as sense$_b$ to be any different to the cost of misclassifying an instance of sense$_b$ as sense$_a$.[1] However, most words have many more than two senses; Zipf (1945) found the most commonly used words tend to have a much greater degree of polysemy than infrequently used words. While accuracy provides a good measure for comparison (satisfying criterion 1) and is simple to comprehend (4), it does not account for non-uniform classification costs (2), meaning that the ranking given will often not reflect the real costs of errors.

2.1 Precision and recall

These problems with accuracy led to the adoption of precision and recall instead of (or in addition to) accuracy for performance measurement. The combination of precision and recall have been used as the primary means of performance evaluation in the SENSEVAL exercises.

Precision and recall are commonly used metrics in information retrieval (IR) (Baeza-Yates and Ribeiro-Neto, 1999). The retrieval task often involves finding a small number of relevant documents from a large data repository. Algorithms are ranked based on their precision/recall tradeoff; an algorithm can be said to be better than another if it has higher precision (recall) for the

[1] This may not be true for all WSD tasks.

same or higher recall (precision). This provides only a loose ranking capacity (criterion 1).

Precision by itself is not a highly relevant measure in WSD as it focuses solely on the positive classifications, treating the negative instances as junk. Unlike IR classification, when disambiguating two senses of an ambiguous word, the set of positives is equally important as the set of negatives, since each corresponds to a distinct sense. The classification question could just as easily be phrased in the negative – this should not affect the performance measure. While high recall on its own would constitute a passable WSD method (in that the set of positive instances are largely correctly classified), high precision alone does not say much about the performance of the method. Simply selecting a single correct positive instance will yield the best possible precision, however, this method will perform woefully.[2] Similarly, classifying all instances as positive will achieve a recall of 1.0 and a precision of $\Pr(P)$ – the proportion of positive instances. As with predictive accuracy, the precision would need to be interpreted with respect to the baseline performance to allow comparisons between different tasks (hence having issues with criterion 3).

When extended to classification of three or more senses, these measures falter. In the case of SENSEVAL, the precision is redefined as the proportion of correctly predicted senses within the set of instances for which the algorithm hazarded a prediction, and recall as the proportion of correctly predicted senses over all instances. This implicitly allows classifiers to opt not to classify every instance. However non-exhaustive classifiers are of limited use, given that they must be combined with other classifiers in order to fully disambiguate a given text. Many tasks in which WSD forms a sub-task, such as machine translation (MT), require the word to be fully disambiguated – an unknown value is unacceptable.

Plotting the precision-recall curves (Manning and Schutze, 2000) allows for better performance ranking by optimising precision for a given level of recall. This goes some way in addressing the issues when assessing precision and recall with respect to criterion (1), however the problem exists as to what recall limit is acceptable – there is no theoretical justification for choosing a specific value, and modifying the value may well alter the rankings of the classifiers. The F-measure (a harmonic mean between precision and recall), may be used for simpler ranking providing a single number for comparison (4). However the weighting assigned to precision and recall in the calculation of the mean needs to be chosen and again, theory does not suggest what values to use.

Criterion (2) is not satisfied by this evaluation metric. The precision and recall values for disambiguation tasks involving three or more senses are based on the number of correct responses, ignoring the types of misclassification. Hence this method suffers for the same problems of predictive accuracy in this regard. Combining precision and recall measured for a number of binary disambiguation tasks for a single word (either between every pairing of senses or between each sense and all other senses) may go some way to satisfying (2) while remaining sensitive to the misclassification costs.

2.2 Semantic/communicative distance

Due to the insensitivity of accuracy and precision and recall to non-uniform misclassification costs, Resnik and Yarowsky (1999) proposed a metric incorporating the costs by weighting misclassification penalties by the distances between the predicted and correct senses. In such a manner misclassifications between fine-grained senses (eg., polysemy) will be penalised less harshly than those between coarser sense distinctions (eg., homonymy). They describe a sense hierarchy for the word *bank* derived from a single or multiple dictionaries, from which they derive a matrix of semantic distance between the senses.

The definition of a sense is a contentious issue within the field. The required granularity of sense distinctions varies with the task in which WSD is used. IR and speech synthesis require only coarse sense distinctions, however for MT and full text understanding much finer distinctions are required – often finer than offered by monolingual dictionaries. This would mean that the set of senses and the misclassification costs between senses, as approximated by the semantic distance, will be task dependent.

In most sense-tagged corpora, sense definitions

[2] Note also that selecting nothing will not yield a precision value at all, due to a division by zero.

have been taken from dictionary meanings or thesaurus categories. Granularity aside, these definitions have been criticised for the level of disagreement between lexicographers themselves (Kilgarriff, 1997). These result in markedly different descriptions of senses in different dictionaries, with no one dictionary offering a definitive set of sense description or more formal representation than all others. There is no reliable method of combining dictionary senses to reflect the level of granularity required by the task.

Resnik and Yarowsky went on to analyse the translation of different senses of a sample of ambiguous English words into 12 target languages. From this they estimated the probability of the senses being lexicalised differently in the translation into the target language. They found that between 52% (fine-grained polysemy) and 95% (homonymy) of senses were lexicalised differently on average in the target languages. They used these statistics to generate semantic distances between senses, reflecting the likelihood that the sense will have a different translation.

In such a scoring model the ranking of classifiers is highly sensitive to the sense hierarchy definition and its use in creating the distance matrix. If either of these were to change – and given the widespread disagreement between lexicographers with regard to sense definitions, this is highly possible – the set of classifiers would need to be reranked. Even when using the translation based measure of semantic distance, the use of a different set of target languages would be likely to affect the scoring. This has the potential to cause previously non-optimal classifiers to be re-ranked as optimal.

The semantic/communicative distance measure improves on the accuracy measure in that it accounts for non-uniform misclassification costs (2), while still providing a ranking measure (1). Translation based semantic distance measures sidestep a number of the issues involved with the use of dictionary sense inventories but are not without problems. The method still requires normalisation with the baseline performance (3), although the kappa statistic could also be used here. What is lost is simplicity (4) – the score assigned is not readily interpretable, as it is based on the distance matrix, an artificial construct based on unfounded assumptions.

3 ROC, an alternative metric

Receiver Operation Characteristic (ROC) graphs are an evaluation technique born in the field of signal detection which have become *de rigueur* in machine learning in recent years (Provost and Fawcett, 1997; Provost and Fawcett, 2001). A ROC graph plots the tradeoff between true positive rate and false negative rate in a binary classifier as a threshold value is modified. The true positive rate (TPR, or recall) is defined as the proportion of positive instances predicted as positive. The false positive rate (FPR, or fallout) is defined as the proportion of negative instances predicted as positive. The rationale behind graphing the relationship between these two factors for a given classifier is that various uses of the classifier may demand different optimisation criteria – such as maximising the TPR given a highest acceptable FPR, or finding the optimal classifier given the costs of errors and class distribution.

Provost and Fawcett described an algorithm for creating a ROC curve for a binary classifier and introduce the ROC convex hull (ROCCH), a method for determining the set of potentially optimal classifiers regardless of the misclassification costs and class distributions. Srinivasan (1999) extended ROC analysis to deal with non-binary classifiers, representing the rate by which each class is traded off for another class as each axis of ROC space. This leads to $c^2 - c$ dimensional ROC space, where c is the number of classes. The ROCCH can be calculated in $O(n^c)$ time, where n is the number of points in ROC space.

The sheer difficultly of visualising such high dimensional space prompted Fawcett to develop an alternative process. The area under the ROC curve (AUC) represents the probability that a binary classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. This assigns a high score to those classifiers which form the majority of the ROCCH, or are consistently close to the hull. Fawcett (2001) extended AUC to cater for multiple classes by treating a c-dimensional classifier as c binary classifiers (each performing a one-vs-all classification), giving:

$$AUC_{total} = \sum_{i} AUC(c_i) \cdot \Pr(c_i)$$

where $\Pr(c_i)$ is the prior probability of the i-th sense.

WSD performance can be measured by the AUC metric, or by comparing a number of classifiers' performance curves in ROC space. Where the misclassification costs are known, the optimal classifier can be found simply by finding the point on the ROCCH with the lowest cost. The cost is simply the sum of the penalties assigned to incorrect classifications, which may be calculated from the semantic/communicative distances between senses as:

$$\sum_i \Pr(c_i) \sum_j r_{ij} d_{ij}$$

where r_{ij} is the proportion of instances of sense i classified as sense j, and d_{ij} is the distance between senses i and j, which is zero when $i = j$.

Where the misclassification costs are unknown or are not known precisely (as would be the case if Resnik and Yarowsky's was supplemented with confidence ranges for each cost), the ROCCH allows performance comparison between the different classifiers. The optimal sub-surface of the ROCCH can be found using the misclassification cost ranges meaning that only classifiers forming part of this sub-surface can be optimal. When the sub-surface is sufficiently small (i.e. the misclassification costs are known to a high degree of confidence) this should provide a good ranking of classifiers, as only a small number will form part of the optimal surface. This allows optimisation of learning methods that cannot incorporate non-uniform misclassification costs, as well as allowing optimisation where these costs are only known approximately and thus cannot be easily incorporated into classifier training. Storing the ROCCH allows this approach to be repeated if misclassification costs were to change.

When the sub-surface is quite large (i.e. when misclassification costs are not known precisely), it is likely that a number of classifiers will lie on the optimal surface. The AUC could then be used to discriminate between these classifiers, ranking those classifiers which are consistently closer to the ROCCH higher than those which are not. While the AUC doesn't strictly indicate optimality, it does provide a reasonable approximation.

This method allows comparison and loose ranking of classifiers (criterion 1), in that a number of classifiers can be discarded. Given precise misclassification costs (2), the classifiers (and indeed combinations of classifiers) can be readily ranked. The baseline performance is implicitly used in the analysis: only those classifiers which achieve better results than (weighted) random combinations of the trivial classifiers will be considered (3). This method has the added benefit of being robust in the face of changing or imprecise misclassification costs. While it does not provide a readily interpretable measure (4), especially when considering the convex hull in high dimensional space, the AUC can provide such a measure.

4 Empirical results and discussion

I have implemented three supervised WSD methods and analysed their performance using the three measures described above. All development was performed in the Natural Language Toolkit (Loper and Bird, 2002) and the source code is available as part of the toolkit. I implemented Yarowsky's (1994) decision list method, which he used for accent restoration in French and Spanish text (roughly similar to homograph disambiguation). This method uses the single most reliable piece of evidence in predicting the sense. I also implemented Brown et al.'s (1991) method, which was used for MT between French and English using decision trees to resolve the correct translation of each ambiguous word. Training uses the *flip-flop* algorithm (Nadas et al., 1991) to determine which feature will maximise the mutual information between a binary division of the values for that feature and the set of most probable senses given the feature takes one of those values. Both of these methods used collocates in a small window around the word as features. Lastly, I created a naive Bayes classifier (Manning and Schutze, 2000), using the unordered bag of words around the ambiguous word as the feature space. Words occurring fewer than five times in the corpora were ignored.

The three algorithms were compared on the interest corpus (Bruce and Wiebe, 1994). The word *interest* has six senses in the corpus with differing degrees of similarity to each other. Four experiments were performed; the first involved disam-

sense of *interest*	f	test$_1$	test$_2$	test$_3$	test$_4$
give attention	15%	✓	✓		✓
worthy of attention	1%	✓			✓
receiving attention	3%				✓
advantage	8%		✓		✓
share of company	21%			✓	✓
money	53%			✓	✓
baseline		97%	85%	72%	52%

Table 1: Test descriptions and baselines.

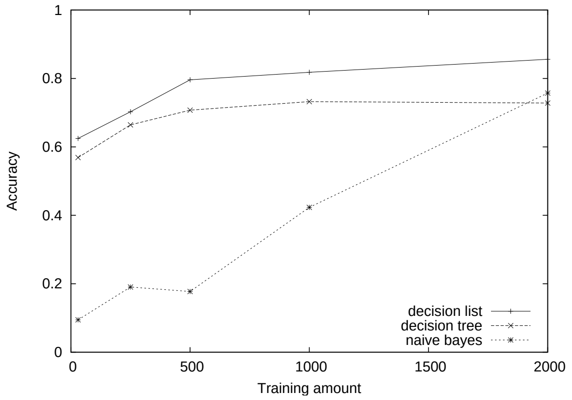


Figure 1: Learning curves

biguating between a pair of fine senses, which reported were difficult for human annotators (Bruce and Wiebe, 1998), and the second and third involve pairs of more distinct senses. The last test involved disambiguating between all six senses. Table 1 shows the gloss for each sense and the senses used for each test.

The learning curve, show in Figure 1, was constructed (in the same vein as Mooney's (1996) performance survey), showing the accuracy of each method on test 4 when trained with increasing amounts of data. It shows all three methods improving, with only the decision tree method showing signs of over-fitting. The accuracy, precision, recall and AUC values were measured and are shown in Table 2. Each test was performed using 10-fold cross validation. The precision, recall and AUC values were calculated with respect to the minority sense for tests 1 - 3. In test 4 both precision and recall are equal to the accuracy, as all three classifiers predict a sense for every instance. ROC curves were generated by ranking each instance (and predicted classification) in order of confidence, using the method described by Provost and Fawcett (2001), from which the AUC measures were calculated. The ROC curves for tests 1 - 3 are shown in Figure 2.

The decision list classifier is shown to be significantly more accurate than the other classifiers, exceeding the baselines for all tests, and performing extremely well for test 3. The results for test 1 are interesting in that the decision list method manages to outperform the baseline performance of 97%. With so few instances no solid conclusions may be drawn, however, the high AUC for the decision tree method suggests that it would perform better (in terms of predictive accuracy) by adjusting its threshold. This would allow it to operate at a more suitable point on its ROC curve, rather than at the origin.

The increase in performance of all methods from test 2 to 3 is most likely due to the increase in data. There are roughly three times as many instances in test 3, providing more training examples. Otherwise, the problems are quite similar, with similar ratios between the two senses. The AUC values support these conclusions, with the decision list and decision tree consistently outperforming naive Bayes for the first three tests. This can also be seen in the ROC curves (Figure 2), where these two classifiers largely dominate naive Bayes. Naive Bayes has a quite low AUC on all of the tests, while still being greater than the benchmark of 0.5. This is reflected in its lower accuracy in each test, however, in test 4, it outperforms the decision tree method despite having a much lower AUC. This suggests that the naive Bayes classifier is operating closer to the point which maximises accuracy on its ROC surface, whereas the decision tree is not. As earlier, this result suggests that the decision tree classifier should be operating with a lower threshold to achieve a higher accuracy. This is also evident in Figure 2, where the curve for the decision tree method, while largely dominated by the decision list curve, is still quite close to the ROCCH.

The highest accuracy classifier would fall on the ROC convex hull at a very steep gradient, due to the minority sense being treated as positive ($m = \frac{TPR}{FPR} = \frac{\Pr(s_b)}{\Pr(s_a)}$ where s_a and s_b are the minority and majority senses respectively). If misclassification costs were biased in favour of the minority sense, the difference in performance between the decision list and decision tree methods would be likely to be reduced, as can be seen from

	test_1	test_2	test_3	test_4
DL - accuracy	97.8	89.1	96.4	85.7
precision	0.8	31.1	26.5	85.7
recall	27.8	83.5	89.1	85.7
AUC	78.1	91.9	95.1	95.6
DT - accuracy	97.0	85.2	95.1	72.0
precision	0.0	27.9	25.4	72.0
recall	0.0	72.8	84.1	72.0
AUC	89.3	83.7	88.5	91.1
NB - accuracy	65.6	78.1	94.3	76.2
precision	3.3	37.1	26.3	76.2
recall	83.3	89.0	86.8	76.2
AUC	53.1	67.4	67.6	60.0

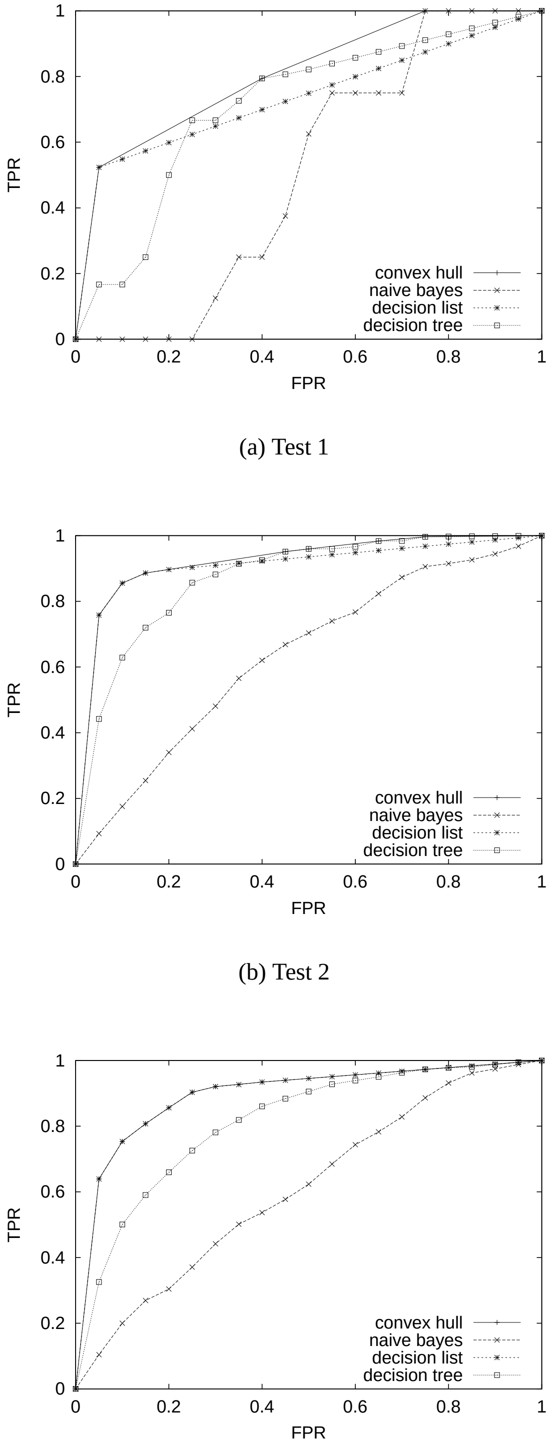
Table 2: Results expressed as percentages.

the proximity of their ROC curves at low gradients. The decision list classifier is shown to be superior to the other two, with higher AUC values on most tests and can be seen to be largely dominating the ROCCH for test 2 and test 3. If the misclassification costs are known at the time of training, a number of learning methods (i.e. naive Bayes) can incorporate them into the training phase, optimising the classifier with respect to these costs. However, this is not possible for all classifiers, requiring the use of ROC analysis to select the optimal classifier.

While the accuracy, precision and recall measures are relatively useful for analysing tests 1 - 3 (assuming uniform misclassification costs), they are not very useful in test_4. The manner in which they aggregate the set of incorrect classifications together loses a great deal of information about the classifier performance. The additional effort required in performing ROC analysis is well rewarded, with much more informative measures of performance.

5 Conclusion

The nebulous nature of the word sense along with differing lexicographic practices mean that the task of WSD is ill-defined. Both dictionary and corpus based definitions of word senses, while not always agreeing on sets of senses for a given word, do concur that some sense pairs are more closely related than others. These relationships have been quantified in deriving the semantic/communicative distance matrix.



(a) Test 1

(b) Test 2

(c) Test 3

Figure 2: ROC curves for tests 1 – 3

ROC analysis proves to be a viable method for analysing performance, addressing a number of shortcomings with the existing measures. It has been shown to be of particular value in measuring performance when disambiguating between three or more senses. It satisfies the objectives of ease of comparison (1), taking misclassification costs into account (2) and implicitly incorporates baseline performance (3), while providing a simple and understandable measure (4) through the AUC. It has the added benefit of being flexible in the face of changing or imprecise misclassification costs. This is of particular significance in WSD given the vigour of the debate over what constitutes a sense, and as to how senses relate to each other. However, ROC analysis suffers from complexity in the form of high dimensional ROC space and computational demands in finding the convex hull.

SENSEVAL, and indeed the whole WSD field, stand to benefit from using ROC analysis as a performance metric. Further research into ROC analysis and its application to WSD and other natural language processing tasks can only help the field mature.

References

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison Wesley.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1991. Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Meeting of the Association for Computational Linguistics*, pages 264–270.

Rebecca Bruce and Janyce Wiebe. 1994. Word sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 139–145, Las Cruces, US.

Rebecca Bruce and Janyce Wiebe. 1998. Word sense distinguishability and inter-coder agreement. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP-98)*, Granada, Spain, June. Association for Computational Linguistics.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: An overview. In *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5.

Tom Fawcett. 2001. Using rule sets to maximize ROC performance. In *2001 IEEE International Conference on Data Mining*, pages 131–138.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Nancy Ide and Jean Veronis. 1998. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):140.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Adam Kilgarriff. 1998. SENSEVAL: An exercise in evaluating word sense disambiguation programs. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 581–588, Granada, Spain.

Sidney I. Landau. 2001. *Dictionaries: The Art and Craft of Lexicography*. Cambridge University Press, second edition.

Claudia Leacock, Geoffrey Towell, and Ellen Voorhees. 1993. Towards building contextual representations of word senses using statistical models. In *SIGLEX workshop: Acquisition of Lexical Knowledge from Text, ACL.*

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, July. Association for Computational Linguistics.

Christopher D. Manning and Hinrich Schutze. 2000. *Foundations of Statistical Natural Language Processing*. MIT Press.

Raymond J. Mooney. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In Eric Brill and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 82–91. Association for Computational Linguistics, Somerset, New Jersey.

Arthur Nadas, David Nahamoo, Michael A. Picheny, and Jeffrey Powell. 1991. An iterative flip-flop approximation of the most informative split in the construction of decision trees. In *International Conference on Acoustics, Speech, and Signal Processing*, New York.

Foster J. Provost and Tom Fawcett. 1997. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Third International Conference on Knowledge Discovery and Data Mining*, pages 43–48.

Foster J. Provost and Tom Fawcett. 2001. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–134.

Ashwin Srinivasan. 1999. Note on the location of optimal classifiers in n-dimensional ROC space. Technical Report PRG-TR-2-99, Oxford University Computing Laboratory, Oxford.

David Yarowsky. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 88–95.

George Zipf. 1945. The meaning-frequency relationship of words. In *Journal of General Psychology*, volume 3, pages 251–256.