Error Bounds for Context Reduction and Feature Omission

Eugen Beck[1]*, Ralf Schlüter*[1]*, Hermann Ney*[1,2]

[1]Human Language Technology and Pattern Recognition, Computer Science Department
RWTH Aachen University, Ahornstr. 55, 52056 Aachen, Germany
[2] Spoken Language Processing Group, LIMSI CNRS, Paris, France
{beck, schlueter, ney}@cs.rwth-aachen.de

Abstract

In language processing applications like speech recognition, printed/handwritten character recognition, or statistical machine translation, the language model usually has a major influence on the performance, by introducing context. An increase of context length usually improves perplexity and increases the accuracy of a classifier using such a language model. In this work, the effect of context reduction, i.e. the accuracy difference between a context sensitive, and a context-insensitive classifier is considered. Context reduction is shown to be related to feature omission in the case of single symbol classification. Therefore, the simplest non-trivial case of feature omission will be analyzed by comparing a feature-aware classifier that uses an emission model to a prior-only classifier that statically infers the prior maximizing class only and thus ignores the observation underlying the classification problem. Upper and lower tight bounds are presented for the accuracy difference of these model classifiers. The corresponding analytic proofs, though not presented here, were supported by an extensive simulation analysis of the problem, which gave empirical estimates of the accuracy difference bounds. Further, it is shown that the same bounds, though not tightly, also apply to the original case of context reduction. This result is supported by further simulation experiments for symbol string classification.
Index Terms: language model, context, error bound

1. Introduction

In applications like automatic speech recognition, statistical machine translation, printed/or handwritten character recognition, classification refers to string classes, where each class represents a string (or sequence) of symbols (words, characters, phonemes, etc.). The corresponding *language models*, providing symbol probability distributions in symbol sequence context of varying length, are an important aspect of many natural language processing tasks. Language modeling paradigms may be based on smoothed n-gram counts [8], or on multilayer perceptrons [2]. Empirically, using longer context improves perplexity and, up to some extent, also the accuracy [13] of string classifiers. Nevertheless, to the best of the authors' knowledge, currently no *formal* relation is known between the order of the *Markov* model used in the language model and the accuracy of a resulting recognition system.

To discover corresponding bounds, an empirical Monte-Carlo approach was applied. To judge if a measure is a potential candidate for a bound, millions of distributions were simulated, discarding measures that did not exhibit a suitable bounding behavior on the accuracy difference of two classifiers with different context length. If a bound existed, its functional form was conjectured, followed by an attempt to find a formal proof.

Information theory provides a number of bounds on the *Bayes* error itself. Examples for this are the *Chernoff* bound [4], the *Lainiotis* bound [10], and the nearest neighbor bound [7]. These bounds do not provide information on the effect of context in string classification, although the nearest-neighbor bound resembles a part of the lower bound presented here. In [5], an upper bound on the *Bayes* error of a string classifier using two classes is described. The bound is a function of the class prior and requires a restriction on the class conditional observation distribution. In [11], two bounds on the accuracy difference between a Bayes single symbol classifier and a model classifier (e.g. one learned from data) are presented. These bounds are based on the squared distance and the Kullback-Leibler divergence [9]. The Kullback-Leibler based bound was later tightened and extended to the general class of f-divergences [6] in [12].

In this work, the feature-dependence of a classifier is analyzed by comparing a feature-aware classifier using an emission model to a prior-only classifier that statically infers the prior maximizing class only. The corresponding accuracy difference between such a pair of classifiers is shown to be closely related to the accuracy difference between a context sensitive, and a context-insensitive classifier, being the original motivation for this work. Upper and lower tight bounds are presented for this accuracy difference. Although not presented here, analytic proofs are available. Extensive simulation analysis of the problem provided the initial hypothesis that lead to these proofs. Further derivations presented here also show that the derived bounds can be related to the accuracy difference induced by context length variation in a language model for symbol string classification, which is supported by simulation results.

2. Context Reduction vs. Feature Omission

Let C be a finite set of classes (e.g. words, symbols, etc.) and X be the set of observations. For simplicity X is assumed to be finite. Then the task of string classification is to map a sequence of observations $x_1^N \in X^N$ to a sequence of classes $c_1^N \in C^N$. Note that here the sequence of classes and observations have the same length and no alignment problem is assumed, like in automatic speech recognition. An exemplary task, which would be represented by this model would be part-of-speech tagging. Let

$$pr(c_1^N, x_1^N) = pr(c_1^N) \cdot pr(x_1^N | c_1^N)$$

be the probability mass function of the true joint distribution, with the language model $pr(c_1^N)$ and the observation model $pr(x_1^N | c_1^N)$. Then the accuracy of a Bayes classifier at position

i in the string of classes is:

$$A_i^* = \sum_{x_1^N} \max_c \left\{ \sum_{c_1^N : c_i = c} pr(c_1^N) pr(x_1^N | c_1^N) \right\}$$

The language model is assumed to be a bigram:

$$pr(c_1^N) = \prod_{n=1}^{N} pr(c_n | c_{n-1})$$

From this bigram a position dependent unigram can be derived by marginalization for position $i \leq N$:

$$pr_i(c) = \sum_{c_1^N : c_i = c} pr(c_1^N) = \sum_{c_1^i : c_i = c} pr(c_1^i)$$

Also, it is assumed that the observation model $pr(x_1^N | c_1^N)$ only exhibits local dependence:

$$pr(x_1^N | c_1^N) = \prod_{n=1}^{N} pr(x_n | c_n)$$

To measure the effect of the language model context, the difference ΔA_i between the full, bigram-based classifier's accuracy A_i^*, and the accuracy of the reduced context classifier $\tilde{A}_i$ that is based on the derived unigram prior, is considered:

$$\Delta A_i = A_i^* - \tilde{A}_i = \sum_{x_1^N} \max_c pr_i(c, x_1^N) - \sum_x \max_c pr_i(c, x),$$

with:

$$pr_i(c, x_1^N) := \sum_{c_1^N : c_i = c} pr(c_1^N, x_1^N),$$
$$pr_i(c, x) := pr_i(c) pr(x|c).$$

To emphasize the connection to single symbols, the last equation is rewritten as follows:

$$\Delta A_i = \sum_{x_i} pr_i(x_i) \Delta A_i(x_i), \tag{1}$$

with the definition of the local accuracy difference:

$$\Delta A_i(x_i) := \sum_{y = x_1^N \setminus x_i} pr_i(y|x_i) \max_c pr_i(c|y, x_i) - \max_c pr_i(c|x_i), \tag{2}$$

and the marginals in symbol position i are, with $y = x_1^N \setminus x_i$:

$$pr_i(x) = \sum_{c_1^N, x_1^N : x_i = x} pr(c_1^N) pr(x_1^N | c_1^N)$$
$$pr_i(c|x) = \frac{pr_i(c) pr(x|c)}{pr_i(x)}$$
$$pr_i(c|y, x_i) = pr_i(c|x_1^N) = \frac{pr_i(c, x_1^N)}{\sum_{c'} pr_i(c', x_1^N)}$$
$$pr_i(c, y|x_i) = pr_i(c, x_1^N \setminus x_i | x_i) = \frac{pr_i(c, x_1^N)}{pr_i(x_i)}$$
$$pr_i(y|x_i) = \sum_c pr_i(c, y|x_i)$$

The local accuracy difference defined in Eq. (2) actually shows the difference between the accuracies of a single symbol classifier that maps an observation $y \in Y$ to a single class $c \in C$, and a classifier that only uses the prior (mapping every observation to the same class). Discarding the condition on x_i and replacing y with x, the accuracy difference for the case of feature omission is obtained:

$$\Delta A = A^* - \tilde{A} = \sum_x \max_c pr(c) pr(x|c) - \max_c pr(c), \tag{3}$$

for which bounds will be derived in the following section that also lead to similar bounds for the symbol string classification case introduced here.

3. Gini Difference Bounds

Assume single symbol classification, and define the following statistical measure for the difference between the class posterior and the class prior probability, which will be called *Gini* difference in the following:

$$\begin{aligned} \Delta G &= \sum_x pr(x) \sum_c pr(c|x)^2 - \sum_c pr(c)^2 \\ &= \sum_x pr(x) \sum_c \left[pr(c|x) - pr(c) \right]^2 \end{aligned}$$

The term *Gini* difference is chosen here, as it is similar to the *Gini* criterion, as, e.g. used in decision tree learning. In [7], the minuend and subtrahend of the *Gini* difference are known as Bayesian distance.

In the following, tight lower and upper bounds of the accuracy difference for the case of feature omission are presented in terms of the *Gini* difference. The corresponding proofs are not presented for lack of space, but are available from the authors on request.

Note that both the *Gini* difference, and the accuracy difference can take values between 0 and $\frac{|C|-1}{|C|}$. Therefore, both measures are be normalized:

$$\Delta A' = \frac{|C|}{|C| - 1} \Delta A,$$
$$\Delta G' = \frac{|C|}{|C| - 1} \Delta G.$$

As shown in the following, in terms of these normalized measures, the bounds do not explicitly depend on the number of classes $|C|$.

3.1. Upper Bound

The normalized accuracy difference defined in Eq. (3) is tightly bounded from above by the square root of the normalized *Gini* difference:

$$\Delta A' \leq \sqrt{\Delta G'}.$$

3.2. Lower Bound

The lower bound of the *Gini* difference consists of three different segments.

3.2.1. First Segment of the Lower Bound

The (normalized) accuracy difference is positive:

$$\Delta A' \geq 0, \tag{4}$$

and equality can be obtained iff the normalized *Gini* difference is constrained to:

$$0 \leq \Delta G' \leq \frac{1}{4}.$$

3.2.2. Second Segment of the Lower Bound

Also, the normalized accuracy difference is linearly bounded from below by the normalized *Gini* difference minus a constant:

$$\Delta A' \geq \Delta G' - \frac{1}{4}$$

This bound is tight for $\frac{1}{4} \leq \Delta G' \leq \frac{3}{4}$.

3.2.3. Third Segment of the Lower Bound

If the *Gini* difference is constrained to

$$\Delta G' \geq \frac{3}{4}, \tag{5}$$

then the set of tight lower bounds of the normalized accuracy difference is completed by:

$$\begin{aligned} &\Delta A' \geq 1 - \sqrt{1 - \Delta G'} \\ \Leftrightarrow\ &\Delta G' \leq 2\Delta A' - (\Delta A')^2 \end{aligned}$$

The bounds are shown in Fig. 1 in terms of normalized *Gini* difference and normalized accuracy difference.

3.3. Transition to Context Reduction

For the case of symbol string classification, the *Gini* difference can also be defined for a specific symbol position i:

$$\Delta G_i := \sum_{x_i} pr_i(x_i) \Delta G_i(x_i)$$

with the local *Gini* difference:

$$\Delta G_i(x_i) := \sum_{y=x_1^N \setminus x_i} pr_i(y|x_i) \sum_c pr_i(c|y, x_i)^2 - \sum_c pr_i(c|x_i)^2$$

Apart from the additional condition on x_i, both the local accuracy difference $\Delta A_i(x_i)$, and the local *Gini* difference $\Delta G_i(x_i)$ effectively can be identified as single symbol cases, such that the same upper and lower bounds apply, as derived for the feature omission case in Subsecs. 3.1, and 3.2. Also, note that these upper and lower bounds are concave and convex functions, respectively. Now assume, these upper and lower bounds are represented by the following two functions g and f, respectively (now assumed without normalization of *Gini*, and accuracy difference, without loss of generality), such that:

$$\Delta A_i(x_i) \leq g\big(\Delta G_i(x_i)\big) \tag{6}$$

$$\Delta A_i(x_i) \geq f\big(\Delta G_i(x_i)\big) \tag{7}$$

Then *Jensen*'s inequality [3, p. 182] can be applied to obtain the same bounds for the global, symbol string case:

$$\begin{aligned} \Delta A_i &= \sum_{x_i} pr_i(x_i) \Delta A_i(x_i) \\ &\leq \sum_{x_i} pr_i(x_i) g\big(\Delta G_i(x_i)\big) \quad \text{(Eq.(6))} \\ &\leq g\big(\sum_{x_i} pr_i(x_i) \Delta G_i(x_i)\big) \quad \text{(\textit{Jensen}'s ineq., concave case)} \\ &\leq g\big(\Delta G_i\big) \\ \Delta A_i &= \sum_{x_i} pr_i(x_i) \Delta A_i(x_i) \\ &\geq \sum_{x_i} pr_i(x_i) f\big(\Delta G_i(x_i)\big) \quad \text{(Eq.(7))} \\ &\geq f\big(\sum_{x_i} pr_i(x_i) \Delta G_i(x_i)\big) \quad \text{(\textit{Jensen}'s ineq., convex case)} \\ &\geq f\big(\Delta G_i\big) \end{aligned}$$

Nevertheless, it should be mentioned that these global bounds for the symbol string case are not necessarily tight anymore, as is confirmed by the simulations shown in the following section.

4. Simulations

4.1. Feature Omission: Single Symbol Case

In order to determine the exact relation between the *Gini* difference and the accuracy difference, originally millions of distributions were simulated to calculate their values of the *Gini*, and the accuracy difference for a number of configurations. In Fig. 1, the results of such a simulation for 8 classes and a set of 16 different discrete observations is presented. An upper and a lower bound for the accuracy difference as a function of the *Gini* difference is visible. This type of simulation also was performed for other combinations of $|C|$ and $|X|$ and from these results the upper and lower bounds presented in Sec. 3 were hypothesized empirically by extensive analysis of the simulations, which further led to corresponding proofs as presented in [1].

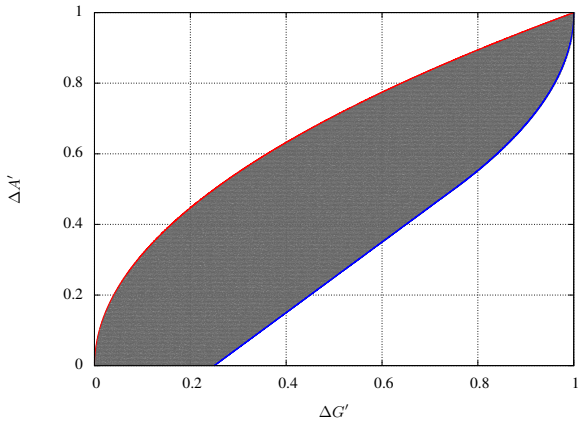

Figure 1: Simulation results for $|C| = 8$ classes and $|X| = 16$ observations. Each gray dot represents one simulated distribution. Also, the derived analytic tight upper and lower bounds are shown in red and blue, respectively.

4.2. Context Reduction: Symbol String Case

The same experiments were performed for symbol string classification. The upper and lower bounds from the symbol case (feature omission) do hold for the string case as shown in Section 2, but the simulations suggest that in this case the bounds are not tight any more, i.e. the simulations do not reach the bound in general, as shown in Fig. 2

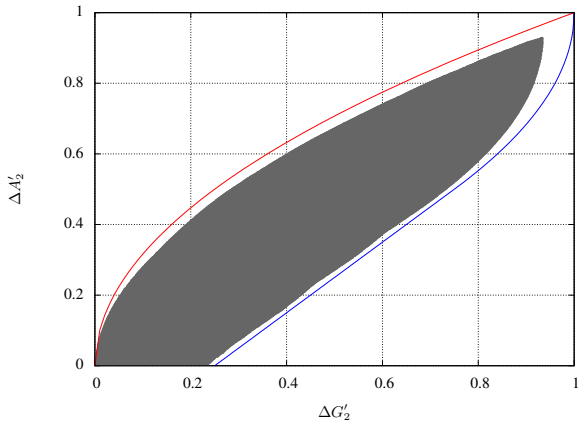

Figure 2: Simulation results for a string classifier with $|C| = 5$ classes, $|X| = 10$ observations, and sequence length $N = 3$. The accuracy/*Gini* difference was calculated at position $i = 2$. Each gray dot represents one simulated distribution.

In the following Fig. 3, the number of classes $|C|$ and observations $|X|$ were proportionally reduced, upon which the space between the analytical bounds is much less filled. This might be due to the dependency between the individual position's distributions, which might be stronger for a lower number of classes and observations.

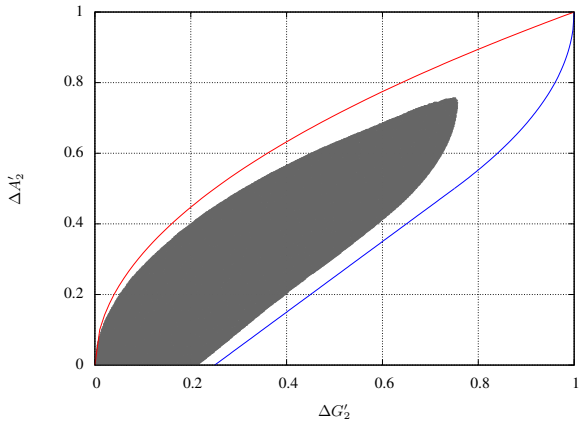

Figure 3: Simulation results for a string classifier with $|C| = 3$ classes, $|X| = 6$ observations, and sequence length $N = 3$. The accuracy/*Gini* difference was calculated at position $i = 2$. Each gray dot represents one simulated distribution.

When (slightly) increasing the length N, apparently no strong difference can be observed, as shown in Fig. 4. The number of observations here was reduced somewhat, as the complexity of the simulations apparently is exponential and the number of simulations required to obtain good filling of the space between the bounds increases strongly.

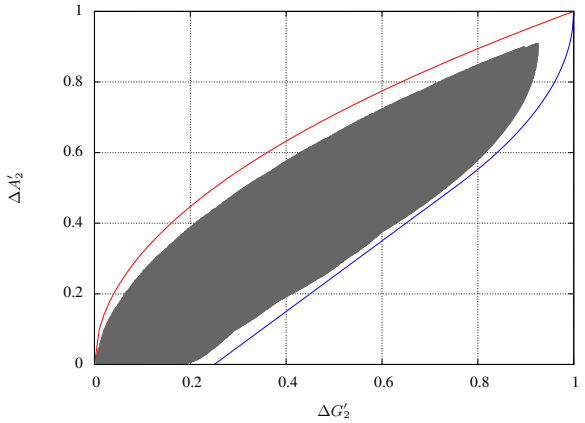

Figure 4: Simulation results for a string classifier with $|C| = 8$ classes, $|X| = 9$ observations, and sequence length $N = 5$. The accuracy/*Gini* difference was calculated at position $i = 3$. Each gray dot represents one simulated distribution.

5. Conclusions & Outlook

In this work, upper and lower bounds on the accuracy difference for feature omission for single symbol classification, and context reduction for symbol string recognition were investigated. First of all, a relation between both cases was derived. Further, tight upper and lower bounds were presented for the single symbol case. Monte-Carlo simulations played an important role in the discovery, as well as the formal proof of the bounds presented. Further simulations for the case of context reduction in symbol string classification were presented, which underline the relation between both cases. As suggested by these, the presented bounds, although being tight for the single symbol case, do not seem to be tight in general for the symbol string case. Nevertheless, the simulations strongly hint at the existence of tighter bounds for the symbol string case, which will be investigated in further work. To the knowledge of the authors, the bounds presented are the first to analytically support the empirically observed effect of feature omission and context reduction on the accuracy.

6. Acknowledgments

The authors would like to thank Tamer Alkhouli and Malte Nuhn for many insightful conversations on this topic. This work has been supported by a compute time grant on the RWTH ITC cluster. This work was partly funded under the project EU-Bridge (FP7-287658). H. Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

7. References

[1] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, Vol. 3, pp. 1137–1155, 2003.

[2] G. Casella, R.L. Berger. *Statistical Inference*, Duxbury Press, Belmont, California, 1990, 650 pages.

[3] H. Chernoff, "A Measure of Asymptotic Efficiency for Tests of

a Hypothesis Based on the Sum of Observations," *The Annals of Mathematical Statistics*, Vol. 23, No. 4, pp. 493–507, 1952.

[4] J. Chu, "Error Bounds for a Contextual Recognition Procedure," *IEEE Transactions on Computers*, Vol. C-20, No. 10, pp. 1203–1207, Oct 1971.

[5] I. Csiszár, "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten," *Magyar. Tud. Akad. Mat. Kutató Int. Közl*, Vol. 8, pp. 85–108, 1963.

[6] P. A. Devijver, "On a New Class of Bounds on Bayes Risk in Multihypothesis Pattern Recognition," *IEEE Transactions on Computers*, Vol. C-23, No. 1, pp. 70–80, Jan. 1974.

[7] R. Kneser and H. Ney, "Improved Backing-Off for m-gram Language Modeling," in *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 181–184, Detroit, MI, May 1995.

[8] S. Kullback and R. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79–86, 1951.

[9] D. Lainiotis, "A class of upper bounds on probability of error for multihypotheses pattern recognition (corresp.)," *IEEE Transactions on Information Theory*, Vol. 15, No. 6, pp. 730–731, Nov. 1969.

[10] H. Ney, "On the Relationship Between Classification Error Bounds and Training Criteria in Statistical Pattern Recognition," in *Proc. Iberian Conference on Pattern Recognition and Image Analysis*, pp. 636–645, Puerto de Andratx, Spain, Jun. 2003.

[11] R. Schlüter, M. Nußbaum-Thom, E. Beck, T. Alkhouli, and H. Ney, "Novel Tight Classification Error Bounds under Mismatch Conditions Based on f-Divergence," in *Proc. IEEE Information Theory Workshop*, pp. 432–436, Sevilla, Spain, Sep. 2013.

[12] H. Schwenk, "Continuous Space Language Models," *Computer Speech & Language*, Vol. 21, No. 3, pp. 492–518, 2007.