# Analyzing and Predicting Language Model Improvements

**R. Iyer†** **M. Ostendorf†** **M. Meteer‡**
† Electrical and Computer Engr. Dept. ‡ BBN Technologies
Boston University GTE Internetworking

**Statistical $n$-gram language models are traditionally developed using perplexity as a measure of goodness. However, perplexity often demonstrates a poor correlation with recognition improvements, mainly because it fails to account for the acoustic confusability between words and for search errors in a recognizer. In this paper, we study alternatives to perplexity for predicting language model performance, including other global features as well as a new approach that predicts, with a high correlation (0.96), performance differences associated with localized changes in language models given a recognition system. Experiments focus on the problem of augmenting in-domain Switchboard text with out-of-domain text from Wall Street Journal and Broadcast News that differ in both style and content from the in-domain data.**

## 1 Introduction

Statistical $n$-gram language models, which model the probability of different word sequences, are traditionally evaluated using one of two measures: perplexity or speech recognition accuracy. While improving word error rate is the eventual goal for language models, it can be computationally expensive to run recognition experiments to evaluate every modification in a language model. Perplexity, on the other hand, is a convenient tool for model development but is often a poor predictor of recognition improvements; mainly because it fails to take into account the acoustic confusability between words and other search issues in a recognizer. Measures that incorporate acoustic confusability in the perplexity framework are much more difficult to compute because of their dependence on acoustic models. Moreover, perplexity and word error rate give only a high-level indication of goodness; neither are diagnostic of weaknesses of a particular language model.

Predicting recognition improvements from language modeling modifications is problematic in general, but it is even more difficult in the case where sparse in-domain training is augmented with out-of-domain text that may differ in both style and content from the in-domain data. Under these conditions, perplexity alone is a poor predictor of performance. In [1], selective use of out-of-domain data led to improved recognition accuracy on the Switchboard task despite degrading perplexity. Figure 1 illustrates that change in

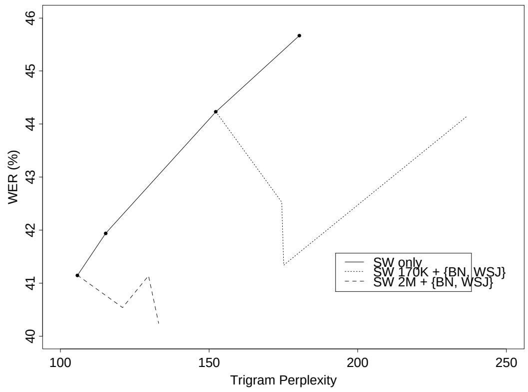


Figure 1: *SW test set perplexity vs. recognition word error rate (WER) for language models that differ in the amount and source of the training data (SW: within domain; BN, WSJ: out of domain).*

perplexity is an indicator of recognition performance changes with increased training data only when that data is within domain.

In this paper, we study alternatives to perplexity for predicting language model performance, including other global features as well as a new approach that predicts performance *differences* associated with localized changes in language models given a recognition system. Experiments focus on the problem of using out-of-domain data.

# 2 Experimental Paradigm

## 2.1 Training

Recognition experiments are conducted on the Switchboard task; transcribing spontaneous, telephone conversational speech. The training data includes 2.1M words of in-domain text from Switchboard (SW), and multi-domain text in the form of 141M words of Broadcast News (BN) data and 40M words of the Wall Street Journal (WSJ) corpus. BN comprises transcribed television text and includes talk shows, discussions, news broadcasts and voice-overs. WSJ is a newspaper text database almost wholly focused on economic and business news. BN is closer in content and style to SW than WSJ.

In order to study the effects of increasing training on recognition performance, we consider several subsets of in-domain and out-of-domain data. The in-domain subsets correspond to 70k, 170k, 1M and 2M words, and the out-of-domain subsets include 141M and 40M words of BN and all of WSJ.

## 2.2 Recognition System

Recognition results are obtained using the BBN Byblos System [2], a speaker-independent HMM system. All language models use the standard trigram framework with the Witten-Bell back-off smoothing. Multi-domain data is combined using a discriminative selection scheme that weights out-of-domain data based on content and style similarity to the given domain [1]. Since we are mainly interested in predicting recognition performance due to changes to the language model, we "fix" all other parameters in the system including: the lexicon (size 10K), the acoustic model (state-clustered tied mixtures), the search paradigm (5-pass search using bigram and trigram models in the first three and last two passes respectively), and the test set.

Table 1 reports recognition results with language models estimated from different amounts of training as described earlier. As expected and as also indicated in Figure 1, recognition performance improves as we increase in-domain training. Adding out-of-domain data further enhances performance, more so with BN than with WSJ but not as much as when adding in-domain data. Not surprisingly, the impact of out-of-domain training data is greater when there is less in-domain data.

In the following sections, we analyze the language models reported in Table 1 at two levels: (i) at the global level considering all words in the given test set, and (ii) at a local per-word level, predicting performance differences associated with a change in language models.

# 3 Global Measures

As pointed out in the introduction, perplexity is not a good indicator of recognition performance unless the data is from the target domain. Therefore, we investigated other measures of language model characteristics in terms of their correlation with word error rate: part-of-speech (POS) class $n$-gram perplexity, $n$-gram coverage, and POS prediction accuracy.

Experiments with POS $n$-gram perplexity showed trends similar to word $n$-gram perplexity: a linear relation to increasing amounts in-domain training and almost no correlation with multi-domain training. Decreasing the order of the $n$-gram did not improve the correlation for either words or POS classes. Interestingly, POS $n$-gram perplexity with BN is consistently lower than with WSJ, further highlighting the style differences between BN and WSJ.

Much better predictability is obtained from $n$-gram coverage, defined as the percentage of words in the test data which have $n$-grams observed at least once in training. As shown in Figure 2, trigram coverage shows a linear relation to WER for in-domain training, similar to perplexity. This relationship extends roughly to the case where out-of-domain data is used, though there is a horizontal shift of the curves associated with the corpus change. POS $n$-gram coverage is not a very useful feature, since POS coverage is very high even with small amounts of training.

Table 1: *SW test set recognition results with different amounts and sources of language model training data.*

| Training | WER % |
|---|---|
| SW 70K | 45.67 |
| SW 170K | 44.23 |
| SW 170K + 40M WSJ | 44.15 |
| SW 170K + 40M BN | 42.53 |
| SW 170K + 141M BN | 41.34 |
| SW 1 M | 41.94 |
| SW 2 M | 41.15 |
| SW 2M + 40M WSJ | 41.14 |
| SW 2M + 40M BN | 40.54 |
| SW 2M + 141M BN | 40.24 |

Owens *et al.* [3] hypothesized that language models that improve POS prediction accuracy will improve WER. Their hypothesis was based on human experiments which showed that, while trigram models perform comparably with humans for predicting a word given the left and right context of words, humans outperform $n$-gram models at predicting the POS class of a word given the same context. While their hypothesis is born out by the in-domain results here, we observe no consistent trend with out-of-domain data.

In summary, we observe that while all three global measures are excellent predictors of performance for increasing in-domain training, only $n$-gram coverage shows some correlation with WER for adding out-of-domain data. However, $n$-gram coverage cannot be used as a general performance predictor since it cannot track improvements due to modifications in the model structure. Once we fix the amount of language training, $n$-gram coverage is fixed. In addition, as for all the global measures, it fails to show that some classes of words improve in accuracy (e.g. content words) while others degrade (e.g. filled pauses and discourse markers). To address the limitations of the $n$-gram coverage statistic and global measures in general, we turn to decision tree analysis techniques based on local per-word features.

# 4 Predicting Performance Differences

Decision trees have previously been used for diagnosing weaknesses of recognition systems [4], where the trees are designed to predict whether or not a word in a test set will be recognized correctly. The prediction is based on features from a single system that is the target of the analysis, and the features may include information associated with the lexicon (e.g. is the word out of vocabulary?), the acoustic model (e.g. how many times has a partic-

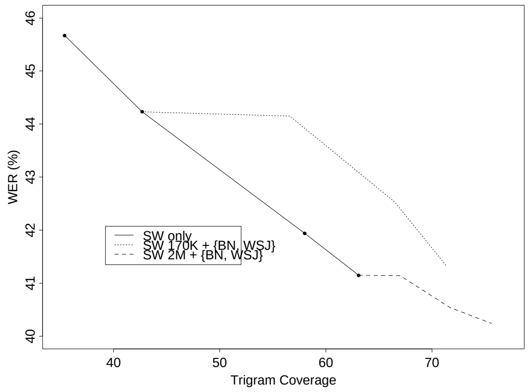


Figure 2: *Word trigram coverage vs. recognition word error rate (WER) on SW with varying amounts and sources of language model training, when the added data is within domain (SW) vs. out of domain (BN/WSJ).*

ular triphone been seen in training?), or the language model (e.g. what is the maximum observed n-gram order for the target word?). The particular features chosen in automatic decision tree design give an indication of the most problematic components of the system. Related rule-based work aimed at error "blame assignment" is described in [5].

In this section, we propose a new tree-based approach aimed at predicting performance *differences* between two systems, rather than analyzing a single system. In addition, the focus is on language modeling. The tree is designed to predict a change in error status of a word between the results of a baseline system and a target new system, and the features aim at capturing differences between the language models of those systems. Neither the baseline nor the target system are assumed fixed, but the acoustic model, dictionary and search parameters are assumed to be the same in both systems. The broad sets of word-level features used include:

- **Delta features:** Comparing scores associated with the target word from the two models under consideration, these features include relative differences in $n$-gram log likelihood, difference in the maximum observed $n$-gram order, and relative differences in the log likelihood of a window of three words centered on the target. The window feature is included since contextual improvements may lead to a better search path.

- **Text-based static features:** These features include POS and content/function/discourse word classifications of the target word, and an indicator of whether the word is in the recognition lexicon.

- **Acoustically-motivated features:** Two features of the target word are included to account for acoustic confusability and search factors:

i) the frequency count of the word in an N-best list generated by the recognition system using the baseline language model (higher frequency indicates less confusability), and ii) the length of the word in phones, based on the hypothesis that longer words are less acoustically confusable with other words.

The leaves of the decision tree perform a three-way classification for every word in the test set: $c_0 = 0$ to represent no change in error, $c_1 = +1$ to indicate improved accuracy, and $c_{-1} = -1$ to represent a degradation in performance, indicating local performance differences between the two models under consideration. Let $C(i)$ be the random variable that represents the change in performance for word $i$, and $E[C(i)]$ is the expected change for that word, which is the sum over the distribution predicted by the tree for that word: $E[C(i)] = \sum_j jc_j(i)$. The predicted change in the percent correct rate for the test set is then $\hat{C} = (1/N)\sum_i E[C(i)]$, where $N$ is the number of words in the test set. The decision tree predictor provides a global metric that predicts the performance gain associated with a new language model, but in addition the choice of questions in the tree indicates what factors led to that gain for analysis purposes.

| Feature | Feature Description |
|---|---|
| F1-F3 | Relative difference in word unigram-trigram likelihood |
| F4 | Length of the word in phones |
| F5 | Relative difference in trigram likelihood of a window of words |
| F6 | Difference in max word $n$-gram order observed in training |
| F7-F8 | Word classification: 15 POS and function/content/discourse |
| F9-F11 | Relative difference in POS unigram-trigram likelihood |
| F12 | Difference in max POS $n$-gram order observed in training |
| F13 | Relative difference in trigram likelihood of a window of POS |
| F14 | Is the word in the recognition vocabulary? |
| F15 | N-best frequency of the word |

Table 2: *Features used for training decision tree for predicting recognition performance differences.*

Decision trees are trained on 15 model pairs (models listed in Table 1), where the members of the pair are estimated from 170K or 2M words of Switchboard and different amounts of BN and WSJ. Each model pair is used in both directions (improved performance and degraded performance) by randomly sampling half of the data for each case, and the tree is trained using some or all of the features listed in Table 2. For a fair evaluation of the decision tree response, we set aside two models, sw170K + 40M WSJ and sw2M + 40M BN, and all corresponding model pairs for testing purposes. However, the other model in the testing pairs is included as one of the models in the training pair.

In Table 3, we report correlation between the predicted and observed performance difference on 13 held-out model pairs (26 considering both directions). The baseline correlation is obtained by predicting performance differences using trigram perplexity differences[1]. Using the word $n$-gram likelihood features, F1-F3, gives a small improvement over using perplexity. Adding F4 as an indicator of acoustic confusability, and F5 as a measure of search improvements further improves correlation. Finally, using all features F1-F15 results in a .96 correlation between the response predicted using decision trees and the observed performance differences. The most important features in the resulting tree are N-best frequency, POS classification and all the likelihood-based features. Figure 3 illustrates the improved performance prediction obtained using the methodology described in this section.

| Training Features | Correlation on Test |
|---|---|
| Perplexity | 0.8300 |
| F1-F3 | 0.8808 |
| F1-F5 | 0.9015 |
| F1-F15 | 0.9653 |

Table 3: *Correlation of decision tree response and actual difference in the percent of correct words for different feature sets, averaging over a test set with language model pairs trained with 170K or 2M SW data plus varying amounts of the out-of-domain data from BN and WSJ.*

# 5 Summary

In summary, we point out the limitations of global measures, particularly perplexity, for predicting recognition performance with new language models when the models are trained with out-of-domain data. We propose a new metric that predicts performance differences between models, using decision trees trained on local per-word features that take into account the two models under consideration and their interactions with other components of the recognition system. Although the approach requires results from several different language models for training the decision trees, it is a powerful performance prediction and analysis tool once developed.

Note that the current results, though quite good, only test the prediction power when the training data and not the model structure is changed. We plan to conduct further experiments to test predictability of performance differences associated with changes in model structure.

[1] Correlation between perplexity and WER in Figure 1 is 0.71.

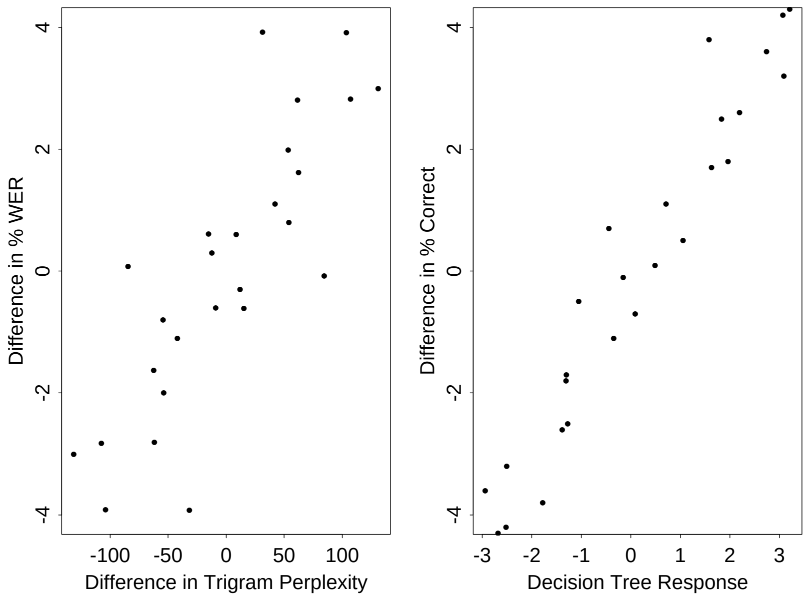


Figure 3: *Comparing two WER predictors: trigram perplexity differences between language models and decision tree response for systems using language models trained with 170K and 2M SW data plus data from BN and WSJ.*

# References


[1] R. Iyer, M. Ostendorf and H. Gish, "Using Out-of-Domain Data to Improve In-Domain Language Models," *IEEE Signal Processing Letters,* **4**(8) 221-223, 1997.

[2] J. Billa *et al.*, "Multilingual Speech Recognition: The 1996 Byblos Callhome System," *Proc. Eurospeech,* **1** 363-367, 1997.

[3] P. O'Boyle and M. A. Owens, "A Comparison Performance with Corpus-Based Language Model Performance on a Task with Limited Context Information," *Third International Conference on the Cognitive Science of Natural Language Processing,* Dublin City University, 149-157, 1994.

[4] E. Eide, *et al.*, "Understanding and Improving Speech Recognition Performance Through the Use of Diagnostic Tools," *Proc. ICASSP*, **I** 221-224, 1995.

[5] L. Chase, "Blame Assignment for Errors Made by Large Vocabulary Speech Recognizers," *Proc. Eurospeech,* **3** 1563-1566, 1997.