An Empirical Investigation of Discounting in Cross-Domain Language Models

Greg Durrett and **Dan Klein**
Computer Science Division
University of California, Berkeley
{gdurrett,klein}@cs.berkeley.edu

Abstract

We investigate the empirical behavior of n-gram discounts within and across domains. When a language model is trained and evaluated on two corpora from exactly the same domain, discounts are roughly constant, matching the assumptions of modified Kneser-Ney LMs. However, when training and test corpora diverge, the empirical discount grows essentially as a linear function of the n-gram count. We adapt a Kneser-Ney language model to incorporate such growing discounts, resulting in perplexity improvements over modified Kneser-Ney and Jelinek-Mercer baselines.

1 Introduction

Discounting, or subtracting from the count of each n-gram, is one of the core aspects of Kneser-Ney language modeling (Kneser and Ney, 1995). For all but the smallest n-gram counts, Kneser-Ney uses a single discount, one that does not grow with the n-gram count, because such constant-discounting was seen in early experiments on held-out data (Church and Gale, 1991). However, due to increasing computational power and corpus sizes, language modeling today presents a different set of challenges than it did 20 years ago. In particular, modeling cross-domain effects has become increasingly more important (Klakow, 2000; Moore and Lewis, 2010), and deployed systems must frequently process data that is out-of-domain from the standpoint of the language model.

In this work, we perform experiments on held-out data to evaluate how discounting behaves in the cross-domain setting. We find that, when training and testing on corpora that are as similar as possible, empirical discounts indeed do not grow with n-gram count, which validates the parametric assumption of Kneser-Ney smoothing. However, when the train and evaluation corpora differ, even slightly, discounts generally exhibit linear growth in the count of the n-gram, with the amount of growth being closely correlated with the corpus divergence. Finally, we build a language model exploiting a parametric form of the growing discount and show perplexity gains of up to 5.4% over modified Kneser-Ney.

2 Discount Analysis

Underlying discounting is the idea that n-grams will occur fewer times in test data than they do in training data. We investigate this quantitatively by conducting experiments similar in spirit to those of Church and Gale (1991). Suppose that we have collected counts on two corpora of the same size, which we will call our train and test corpora. For an n-gram $w = (w_1, ..., w_n)$, let $k_{\text{train}}(w)$ denote the number of occurrences of w in the training corpus, and $k_{\text{test}}(w)$ denote the number of occurrences of w in the test corpus. We define the empirical discount of w to be $d(w) = k_{\text{train}}(w) - k_{\text{test}}(w)$; this will be negative when the n-gram occurs more in the test data than in the training data. Let $W_i = \{w : k_{\text{train}}(w) = i\}$ be the set of n-grams with count i in the training corpus. We define the *average empirical discount* function as

$$\bar{d}(i) = \frac{1}{|W_i|} \sum_{w \in W_i} d(w)$$

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 24–29,
Portland, Oregon, June 19-24, 2011.

Kneser-Ney implicitly makes two assumptions: first, that discounts do not depend on n-gram count, i.e. that $\bar{d}(i)$ is constant in i. Modified Kneser-Ney relaxes this assumption slightly by having independent parameters for 1-count, 2-count, and many-count n-grams, but still assumes that $\bar{d}(i)$ is constant for i greater than two. Second, by using the same discount for all n-grams with a given count, Kneser-Ney assumes that the distribution of $d(w)$ for w in a particular W_i is well-approximated by its mean. In this section, we analyze whether or not the behavior of the average empirical discount function supports these two assumptions. We perform experiments on various subsets of the documents in the English Gigaword corpus, chiefly drawn from New York Times (NYT) and Agence France Presse (AFP).[1]

2.1 Are Discounts Constant?

Similar corpora To begin, we consider the NYT documents from Gigaword for the year 1995. In order to create two corpora that are maximally domain-similar, we randomly assign half of these documents to train and half of them to test, yielding train and test corpora of approximately 50M words each, which we denote by NYT95 and NYT95′. Figure 1 shows the average empirical discounts $\bar{d}(i)$ for trigrams on this pair of corpora. In this setting, we recover the results of Church and Gale (1991) in that discounts are approximately constant for n-gram counts of two or greater.

Divergent corpora In addition to these two corpora, which were produced from a single contiguous batch of documents, we consider testing on corpus pairs with varying degrees of domain difference. We construct additional corpora NYT96, NYT06, AFP95, AFP96, and AFP06, by taking 50M words from documents in the indicated years of NYT and AFP data. We then collect training counts on NYT95 and alternately take each of our five new corpora as the test data. Figure 1 also shows the average empirical discount curves for these train/test pairs. Even within NYT newswire data, we see growing discounts when the train and test corpora are drawn from different years, and between the NYT and AFP newswire, discounts grow even more quickly. We observed these trends continuing steadily up into n-gram counts in the hundreds, beyond which point it becomes difficult to robustly estimate discounts due to fewer n-gram types in this count range.

[1]Gigaword is drawn from six newswire sources and contains both miscellaneous text and complete, contiguous documents, sorted chronologically. Our experiments deal exclusively with the document text, which constitutes the majority of Gigaword and is of higher quality than the miscellaneous text.

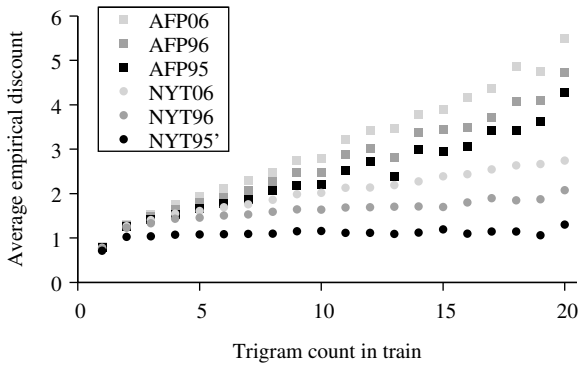

Figure 1: Average empirical trigram discounts $\bar{d}(i)$ for six configurations, training on NYT95 and testing on the indicated corpora. For each n-gram count k, we compute the average number of occurrences in test for all n-grams occurring k times in training data, then report k minus this quantity as the discount. Bigrams and bigram types exhibit similar discount relationships.

This result is surprising in light of the constant discounts observed for the NYT95/NYT95′ pair. Goodman (2001) proposes that discounts arise from document-level "burstiness" in a corpus, because language often repeats itself locally within a document, and Moore and Quirk (2009) suggest that discounting also corrects for quantization error due to estimating a continuous distribution using a discrete maximum likelihood estimator (MLE). Both of these factors are at play in the NYT95/NYT95′ experiment, and yet only a small, constant discount is observed. Our growing discounts must therefore be caused by other, larger-scale phenomena, such as shifts in the subjects of news articles over time or in the style of the writing between newswire sources. The increasing rate of discount growth as the source changes and temporal divergence increases lends credence to this hypothesis.

2.2 Nonuniformity of Discounts

Figure 1 considers discounting in terms of averaged discounts for each count, which tests one assumption of modified Kneser-Ney, that discounts are a

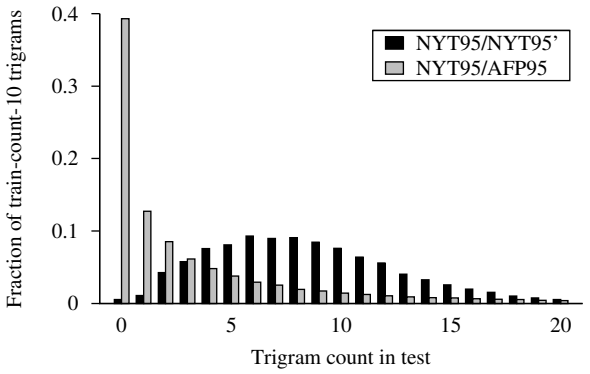

Figure 2: Empirical probability mass functions of occurrences in the test data for trigrams that appeared 10 times in training data. Discounting by a single value is plausible in the case of similar train and test corpora, where the mean of the distribution (8.50) is close to the median (8.0), but not in the case of divergent corpora, where the mean (6.04) and median (1.0) are very different.

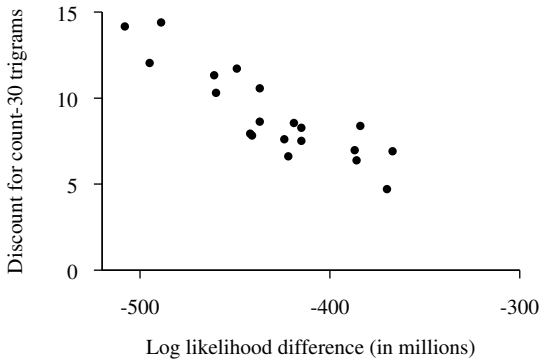

Figure 3: Log likelihood difference versus average empirical discount of trigrams with training count 30 ($\bar{d}(30)$) for the train/test pairs. More negative values of the log likelihood indicate more dissimilar corpora, as the trained model is doing less well relative to the jackknife model.

constant function of n-gram counts. In Figure 2, we investigate the second assumption, namely that the distribution over discounts for a given n-gram count is well-approximated by its mean. For similar corpora, this seems to be true, with a histogram of test counts for trigrams of count 10 that is nearly symmetric. For divergent corpora, the data exhibit high skew: almost 40% of the trigrams simply never appear in the test data, and the distribution has very high standard deviation (17.0) due to a heavy tail (not shown). Using a discount that depends only on the n-gram count is less appropriate in this case.

In combination with the growing discounts of section 2.1, these results point to the fact that modified Kneser-Ney does not faithfully model the discounting in even a mildly cross-domain setting.

2.3 Correlation of Divergence and Discounts

Intuitively, corpora that are more temporally distant within a particular newswire source should perhaps be slightly more distinct, and still a higher degree of divergence should exist between corpora from different newswire sources. From Figure 1, we see that this notion agrees with the relative sizes of the observed discounts. We now ask whether growth in discounts is correlated with train/test dissimilarity in a more quantitative way. For a given pair of corpora, we canonicalize the degree of discounting by selecting the point $\bar{d}(30)$, the average empirical discount for n-grams occurring 30 times in training.[2] To measure divergence between the corpus pair, we compute the difference between the log likelihood of the test corpus under the train corpus language model (using basic Kneser-Ney) and the likelihood of the test corpus under a jackknife language model from the test itself, which holds out and scores each test n-gram in turn. This dissimilarity metric resembles the cross-entropy difference used by Moore and Lewis (2010) to subsample for domain adaptation.

We compute this canonicalization for each of twenty pairs of corpora, with each corpus containing 240M trigram tokens between train and test. The corpus pairs were chosen to span varying numbers of newswire sources and lengths of time in order to capture a wide range of corpus divergences. Our results are plotted in Figure 3. The log likelihood difference and $\bar{d}(30)$ are negatively correlated with a correlation coefficient value of $r = -0.88$, which strongly supports our hypothesis that higher divergence yields higher discounting. One explanation for the remaining variance is that the trigram discount curve depends on the difference between the number of bigram types in the train and test corpora, which can be as large as 10%: observing more bigram contexts in training fragments the token counts

[2] One could also imagine instead canonicalizing the curves by using either the exponent or slope parameters from a fitted power law as in section 3. However, there was sufficient nonlinearity in the average empirical discount curves that neither of these parameters was an accurate proxy for $\bar{d}(i)$.

and leads to smaller observed discounts.

2.4 Related Work

The results of section 2.1 point to a remarkably pervasive phenomenon of growing empirical discounts, except in the case of extremely similar corpora. Growing discounts of this sort were previously suggested by the model of Teh (2006). However, we claim that the discounting phenomenon in our data is fundamentally different from his model's prediction. In the held-out experiments of section 2.1, growing discounts only emerge when one evaluates against a dissimilar held-out corpus, whereas his model would predict discount growth even in NYT95/NYT95′, where we do not observe it.

Adaptation across corpora has also been addressed before. Bellegarda (2004) describes a range of techniques, from interpolation at either the count level or the model level (Bacchiani and Roark, 2003; Bacchiani et al., 2006) to using explicit models of syntax or semantics. Hsu and Glass (2008) employ a log-linear model for multiplicatively discounting n-grams in Kneser-Ney; when they include the log-count of an n-gram as the only feature, they achieve 75% of their overall word error rate reduction, suggesting that predicting discounts based on n-gram count can substantially improve the model. Their work also improves on the second assumption of Kneser-Ney, that of the inadequacy of the average empirical discount as a discount constant, by employing various other features in order to provide other criteria on which to discount n-grams.

Taking a different approach, both Klakow (2000) and Moore and Lewis (2010) use subsampling to select the domain-relevant portion of a large, general corpus given a small in-domain corpus. This can be interpreted as a form of hard discounting, and implicitly models both growing discounts, since frequent n-grams will appear in more of the rejected sentences, and nonuniform discounting over n-grams of each count, since the sentences are chosen according to a likelihood criterion. Although we do not consider this second point in constructing our language model, an advantage of our approach over subsampling is that we use our entire training corpus, and in so doing compromise between minimizing errors from data sparsity and accommodating domain shifts to the extent possible.

3 A Growing Discount Language Model

We now implement and evaluate a language model that incorporates growing discounts.

3.1 Methods

Instead of using a fixed discount for most n-gram counts, as prescribed by modified Kneser-Ney, we discount by an increasing parametric function of the n-gram count. We use a tune set to compute an average empirical discount curve $\bar{d}(i)$, and fit a function of the form $f(x) = a + bx^c$ to this curve using weighted least-L_1-loss regression, with the weight for each point proportional to $i|W_i|$, the total token counts of n-grams occurring that many times in training. To improve the fit of the model, we use dedicated parameters for count-1 and count-2 n-grams as in modified Kneser-Ney, yielding a model with five parameters per n-gram order. We call this model GDLM. We also instantiate this model with c fixed to one, so that the model is strictly linear (GDLM-LIN).

As baselines for comparison, we use basic interpolated Kneser-Ney (KNLM), with one discount parameter per n-gram order, and modified interpolated Kneser-Ney (MKNLM), with three parameters per n-gram order, as described in (Chen and Goodman, 1998). We also compare against Jelinek-Mercer smoothing (JMLM), which interpolates the undiscounted MLEs from every order. According to Chen and Goodman (1998), it is common to use different interpolation weights depending on the history count of an n-gram, since MLEs based on many samples are presumed to be more accurate than those with few samples. We used five history count buckets so that JMLM would have the same number of parameters as GDLM.

All five models are trigram models with type counts at the lower orders and independent discount or interpolation parameters for each order. Parameters for GDLM, MKNLM, and KNLM are initialized based on estimates from $\bar{d}(i)$: the regression thereof for GDLM, and raw discounts for MKNLM and KNLM. The parameters of JMLM are initialized to constants independent of the data. These initializations are all heuristic and not guaranteed to be optimal, so we then iterate through the parameters of each model several times and perform line search

	Train NYT00+01		Train AFP02+05+06	
Voc.	157K	50K	157K	50K
GDLM(*)	151	131	258	209
GDLM-LIN(*)	151	132	259	210
JMLM	165	143	274	221
MKNLM	152	132	273	221
KNLM	159	138	300	241

Table 1: Perplexities of the growing discounts language model (GDLM) and its purely linear variant (GDLM-LIN), which are contributions of this work, versus the modified Kneser-Ney (MKNLM), basic Kneser-Ney (KNLM), and Jelinek-Mercer (JMLM) baselines. We report results for in-domain (NYT00+01) and out-of-domain (AFP02+05+06) training corpora, for two methods of closing the vocabulary.

in each to optimize tune-set perplexity.

For evaluation, we train, tune, and test on three disjoint corpora. We consider two different training sets: one of 110M words of NYT from 2000 and 2001 (NYT00+01), and one of 110M words of AFP from 2002, 2005, and 2006 (AFP02+05+06). In both cases, we compute $\bar{d}(i)$ and tune parameters on 110M words of NYT from 2002 and 2003, and do our final perplexity evaluation on 4M words of NYT from 2004. This gives us both in-domain and out-of-domain results for our new language model. Our tune set is chosen to be large so that we can initialize parameters based on the average empirical discount curve; in practice, one could compute empirical discounts based on a smaller tune set with the counts scaled up proportionately, or simply initialize to constant values.

We use two different methods to handle out-of-vocabulary (OOV) words: one scheme replaces any unigram token occurring fewer than five times in training with an UNK token, yielding a vocabulary of approximately 157K words, and the other scheme only keeps the top 50K words in the vocabulary. The count truncation method has OOV rates of 0.9% and 1.9% in the NYT/NYT and NYT/AFP settings, respectively, and the constant-size vocabulary has OOV rates of 2% and 3.6%.

3.2 Results

Perplexity results are given in Table 1. As expected, for in-domain data, GDLM performs comparably to MKNLM, since the discounts do not grow and so there is little to be gained by choosing a parameterization that permits this. Out-of-domain, our model outperforms MKNLM and JMLM by approximately 5% for both vocabulary sizes. The out-of-domain perplexity values are competitive with those of Rosenfeld (1996), who trained on New York Times data and tested on AP News data under similar conditions, and even more aggressive closing of the vocabulary. Moore and Lewis (2010) achieve lower perplexities, but they use in-domain training data that we do not include in our setting.

We briefly highlight some interesting features of these results. In the small vocabulary cross-domain setting, for GDLM-LIN, we find

$$d_{\text{tri}}(i) = 1.31 + 0.27i,\ d_{\text{bi}}(i) = 1.34 + 0.05i$$

as the trigram and bigram discount functions that minimize tune set perplexity. For GDLM,

$$d_{\text{tri}}(i) = 1.19 + 0.32i^{0.45},\ d_{\text{bi}}(i) = 0.86 + 0.56i^{0.86}$$

In both cases, a growing discount is indeed learned from the tuning procedure, demonstrating the importance of this in our model. Modeling nonlinear discount growth in GDLM yields only a small marginal improvement over the linear discounting model GDLM-LIN, so we prefer GDLM-LIN for its simplicity.

A somewhat surprising result is the strong performance of JMLM relative to MKNLM on the divergent corpus pair. We conjecture that this is because the bucketed parameterization of JMLM gives it the freedom to change interpolation weights with n-gram count, whereas MKNLM has essentially a fixed discount. This suggests that modified Kneser-Ney as it is usually parameterized may be a particularly poor choice in cross-domain settings.

Overall, these results show that the growing discount phenomenon detailed in section 2, beyond simply being present in out-of-domain held-out data, provides the basis for a new discounting scheme that allows us to improve perplexity relative to modified Kneser-Ney and Jelinek-Mercer baselines.

Acknowledgments

The authors gratefully acknowledge partial support from the GALE program via BBN under DARPA contract HR0011-06-C-0022, and from an NSF fellowship for the first author. Thanks to the anonymous reviewers for their insightful comments.

References

Michiel Bacchiani and Brian Roark. 2003. Unsupervised Langauge Model Adaptation. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*.

Michiel Bacchiani, Michael Riley, Brian Roark, and Richard Sproat. 2006. MAP adaptation of stochastic grammars. *Computer Speech & Language*, 20(1):41 – 68.

Jerome R. Bellegarda. 2004. Statistical language model adaptation: review and perspectives. *Speech Communication*, 42:93–108.

Stanley Chen and Joshua Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. Technical report, Harvard University, August.

Kenneth Church and William Gale. 1991. A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams. *Computer Speech & Language*, 5(1):19–54.

Joshua Goodman. 2001. A Bit of Progress in Language Modeling. *Computer Speech & Language*, 15(4):403–434.

Bo-June (Paul) Hsu and James Glass. 2008. N-gram Weighting: Reducing Training Data Mismatch in Cross-Domain Language Model Estimation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 829–838.

Dietrich Klakow. 2000. Selecting articles from the language model training corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1695–1698.

Reinhard Kneser and Hermann Ney. 1995. Improved Backing-off for M-Gram Language Modeling. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, July.

Robert C. Moore and Chris Quirk. 2009. Improved Smoothing for N-gram Language Models Based on Ordinary Counts. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 349–352.

Ronald Rosenfeld. 1996. A Maximum Entropy Approach to Adaptive Statistical Language Modeling. *Computer, Speech & Language*, 10:187–228.

Yee Whye Teh. 2006. A Hierarchical Bayesian Language Model Based On Pitman-Yor Processes. In *Proceedings of ACL*, pages 985–992, Sydney, Australia, July. Association for Computational Linguistics.