Richer Interpolative Smoothing Based on Modified Kneser-Ney Language Modeling

Ehsan Shareghi,[♣] Trevor Cohn[♠] and Gholamreza Haffari[♣]
[♣] Faculty of Information Technology, Monash University
[♠] Computing and Information Systems, The University of Melbourne
first.last@{monash.edu,unimelb.edu.au}

Abstract

In this work we present a generalisation of the Modified Kneser-Ney interpolative smoothing for richer smoothing via additional discount parameters. We provide mathematical underpinning for the estimator of the new discount parameters, and showcase the utility of our rich MKN language models on several European languages. We further explore the interdependency among the training data size, language model order, and number of discount parameters. Our empirical results illustrate that larger number of discount parameters, i) allows for better allocation of mass in the smoothing process, particularly on small data regime where statistical sparsity is severe, and ii) leads to significant reduction in perplexity, particularly for out-of-domain test sets which introduce higher ratio of out-of-vocabulary words.[1]

1 Introduction

Probabilistic language models (LMs) are the core of many natural language processing tasks, such as machine translation and automatic speech recognition. m-gram models, the corner stone of language modeling, decompose the probability of an utterance into conditional probabilities of words given a fixed-length context. Due to sparsity of the events in natural language, smoothing techniques are critical for generalisation beyond the training text when estimating the parameters of m-gram LMs. This is particularly important when the training text is small, e.g. building language models for translation or speech recognition in low-resource languages.

A widely used and successful smoothing method is interpolated Modified Kneser-Ney (MKN) (Chen and Goodman, 1999). This method uses a linear interpolation of higher and lower order m-gram probabilities by preserving probability mass via absolute discounting. In this paper, we extend MKN by introducing additional discount parameters, leading to a richer smoothing scheme. This is particularly important when statistical sparsity is more severe, i.e., in building high-order LMs on small data, or when out-of-domain test sets are used.

Previous research in MKN language modeling, and more generally m-gram models, has mainly dedicated efforts to make them faster and more compact (Stolcke et al., 2011; Heafield, 2011; Shareghi et al., 2015) using advanced data structures such as succinct suffix trees. An exception is Hierarchical Pitman-Yor Process LMs (Teh, 2006a; Teh, 2006b) providing a rich Bayesian smoothing scheme, for which Kneser-Ney smoothing corresponds to an approximate inference method. Inspired by this work, we directly enrich MKN smoothing realising some of the reductions while remaining more efficient in learning and inference.

We provide estimators for our additional discount parameters by extending the discount bounds in MKN. We empirically analyze our enriched MKN LMs on several European languages in in- and out-of-domain settings. The results show that our discounting mechanism significantly improves the perplexity compared to MKN and offers a more elegant

[1] For the implementation see: https://github.com/eehsan/cstlm

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 944–949,
Austin, Texas, November 1-5, 2016.

way of dealing with out-of-vocabulary (OOV) words and domain mismatch.

2 Enriched Modified Kneser-Ney

Interpolative Modified Kneser-Ney (MKN) (Chen and Goodman, 1999) smoothing is widely accepted as a state-of-the-art technique and is implemented in leading LM toolkits, e.g., SRILM (Stolcke, 2002) and KenLM (Heafield, 2011).

MKN uses lower order k-gram probabilities to smooth higher order probabilities. $P(w|\mathbf{u})$ is defined as,

$$\frac{c(\mathbf{u}w) - \mathbb{D}^m(c(\mathbf{u}w))}{c(\mathbf{u})} + \frac{\gamma(\mathbf{u})}{c(\mathbf{u})} \times \bar{P}(w|\pi(\mathbf{u}))$$

where $c(\mathbf{u})$ is the frequency of the pattern $\mathbf{u}$, $\gamma(.)$ is a constant ensuring the distribution sums to one, and $\bar{P}(w|\pi(\mathbf{u}))$ is the smoothed probability computed recursively based on a similar formula[2] conditioned on the suffix of the pattern $\mathbf{u}$ denoted by $\pi(\mathbf{u})$. Of particular interest are the discount parameters $\mathbb{D}^m(.)$ which remove some probability mass from the maximum likelihood estimate for each event which is redistributed over the smoothing distribution. The discounts are estimated as

$$\mathbb{D}^m(i) = \begin{cases} 0, & \text{if } i = 0 \\ 1 - 2\frac{n_2[m]}{n_1[m]}\frac{n_1[m]}{n_1[m]+2n_2[m]}, & \text{if } i = 1 \\ 2 - 3\frac{n_3[m]}{n_2[m]}\frac{n_1[m]}{n_1[m]+2n_2[m]}, & \text{if } i = 2 \\ 3 - 4\frac{n_4[m]}{n_3[m]}\cdot\frac{n_1[m]}{n_1[m]+2n_2[m]}, & \text{if } i \geq 3 \end{cases}$$

where $n_i(m)$ is the number of unique m-grams[3] of frequency i. This effectively leads to three discount parameters $\{\mathbb{D}^m(1), \mathbb{D}^m(2), \mathbb{D}^m(3+)\}$ for the distributions on a particular context length, m.

2.1 Generalised MKN

Ney et al. (1994) characterized the data sparsity using the following empirical inequalities,

$$3n_3[m] < 2n_2[m] < n_1[m] \quad \text{for} \quad m \leq 3$$

It can be shown (see Appendix A) that these empirical inequalities can be extended to higher frequencies and larger contexts $m > 3$,

$$\begin{aligned}(N-m)n_{N-m}[m] < ... < 2n_2[m] \\ < n_1[m] < \sum_{i>0} n_i[m] \ll n_0[m] < \sigma^m\end{aligned}$$

where σ^m is the possible number of m-grams over a vocabulary of size σ, $n_0[m]$ is the number of m-grams that never occurred, and $\sum_{i>0} n_i[m]$ is the number of m-grams observed in the training data.

We use these inequalities to extend the discount depth of MKN, resulting in new discount parameters. The additional discount parameters increase the flexibility of the model in altering a wider range of raw counts, resulting in a more elegant way of assigning the mass in the smoothing process. In our experiments, we set the number of discounts to 10 for all the levels of the hierarchy, (compare this to these in MKN).[4] This results in the following estimators for the discounts,

$$\mathbb{D}^m(i) = \begin{cases} 0, & \text{if } i = 0 \\ i - (i+1)\frac{n_{i+1}[m]}{n_i[m]}\frac{n_1[m]}{n_1[m]+2n_2[m]}, & \text{if } i < 10 \\ 10 - 11\frac{n_{11}[m]}{n_{10}[m]}\cdot\frac{n_1[m]}{n_1[m]+2n_2[m]}, & \text{if } i \geq 10 \end{cases}$$

It can be shown that the above estimators for our discount parameters are derived by maximizing a lower bound on the leave-one-out likelihood of the training set, following (Ney et al., 1994; Chen and Goodman, 1999) (see Appendix B for the proof sketch).

3 Experiments

We compare the effect of using different numbers of discount parameters on perplexity using the Finnish (FI), Spanish (ES), German (DE), English (EN) portions of the Europarl v7 (Koehn, 2005) corpus. For each language we excluded the first 10K sentences and used it as the in-domain test set (denoted as EU), skipped the second 10K sentences, and used the rest as the training set. The data was tokenized, sentence split, and the XML markup discarded. We tested the effect of domain mismatch, under two settings for out-of-domain test sets: i) *mild* using the Spanish section of news-test 2013, the German, English sections of news-test 2014, and the Finnish section

[2]Note that in all but the top layer of the hierarchy, *continuation counts*, which count the number of unique contexts, are used in place of the frequency counts (Chen and Goodman, 1999).

[3]Continuation counts are used for the lower layers.

[4]We have selected the value of 10 arbitrarily; however our approach can be used with larger number of discount parameters, with the caveat that we would need to handle sparse counts in the higher orders.

Training	size (M) tokens	sents	Test	size (K) tokens	sents	OOV%	Perplexity MKN ($D_{1...3}$) $m=2$	$m=5$	$m=10$	MKN ($D_{[1...4]}$) $m=2$	$m=5$	$m=10$	MKN ($D_{[1...10]}$) $m=2$	$m=5$	$m=10$
			NT	19.8	3	9.2	6536.6	5900.3	5897.3	6451.3	5827.6	5824.6	6154.4	5575.0	5572.5
FI	46.5	2.2	EU	197.3	10	6.1	390.7	287.4	286.8	390.7	287.3	286.6	390.4	287.3	286.8
			TW	10.9	1.3	52.1	57 825.1	51 744.1	51 740.1	55 550.2	49 884.2	49 881.3	47 696.2	43 277.3	43 275.5
			NT	70.7	3	9.1	565.6	431.5	429.4	560.0	425.5	423.5	541.5	409.0	407.3
ES	68.0	2.2	EU	281.5	10	2.4	92.7	51.5	51.1	92.8	51.5	51.1	92.8	51.4	51.0
			TW	3141.3	293	78.5	17 804.2	14 062.7	14 027.1	17 121.4	13 487.4	13 454.1	14 915.7	11 832.1	11 807.2
			NT	64.5	3	18.7	2190.7	1784.6	1781.8	2158.9	1755.8	1753.2	2065.3	1680.6	1678.3
DE	61.2	2.3	EU	244.0	10	4.6	156.9	91.7	91.2	156.9	91.6	91.2	156.4	91.7	91.2
			MED	317.7	10	59.8	5135.7	4232.4	4226.7	5007.5	4123.0	4117.5	4636.0	3831.2	3826.6
			NT	69.5	3	5.5	1089.2	875.0	872.2	1071.1	857.2	854.4	1011.5	806.7	804.4
EN	67.5	2.2	EU	274.9	10	1.7	90.1	48.4	48.1	90.1	48.3	48.0	90.5	48.3	48.0
			MED	405.9	10	44.1	2319.7	1947.9	1942.5	2261.6	1893.3	1888.2	2071.9	1734.9	1730.8

Table 1: Perplexity for various m-gram orders $m \in 2, 3, 10$ and training languages from Europarl, using different numbers of discount parameters for MKN. **MKN ($D_{[1...3]}$)**, **MKN ($D_{[1...4]}$)**, **MKN ($D_{[1...10]}$)** represent vanilla MKN, MKN with 1 more discounts, and MKN with 7 more discount parameters, respectively. Test sets sources EU, NT, TW, MED are Europarl, news-test, Twitter, and medical patent descriptions, respectively. OOV is reported as the ratio $\frac{|\{OOV \in \text{test-set}\}|}{|\{w \in \text{test-set}\}|}$.

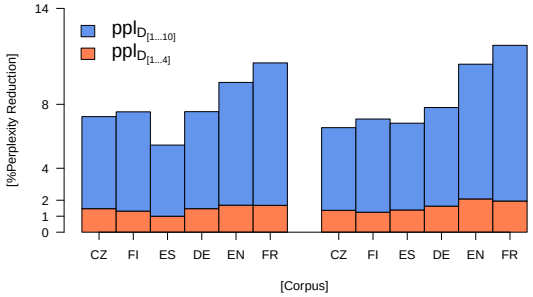

Figure 1: Percentage of perplexity reduction for pplx$_{D_{[1...4]}}$ and pplx$_{D_{[1...10]}}$ compared with pplx$_{D_{[1...3]}}$ on different training corpora (Europarl CZ, FI, ES, DE, EN, FR) and on news-test sets (NT) for $m = 2$ (left), and $m = 10$ (right).

of news-test 2015 (all denoted as NT)[5], and ii) *extreme* using a 24 hour period of streamed Finnish, and Spanish tweets[6] (denoted as TW), and the German and English sections of the patent description of medical translation task[7] (denoted as MED). See Table 1 for statistics of the training and test sets.

3.1 Perplexity

Table 1 shows substantial reduction in perplexity on all languages for out-of-domain test sets when expanding the number of discount parameters from 3 in vanilla MKN to 4 and 10. Consider the English news-test (NT), in which even for a 2-gram language model a single extra discount parameter ($m = 2$, $D_{[1...4]}$) improves the perplexity by 18 points and this improvement quadruples to 77 points when using 10 discounts ($m = 2$, $D_{[1...10]}$). This effect is consistent across the Europarl corpora, and for all LM orders. We observe a substantial improvements even for $m = 10$-gram models (see Figure 1). On the medical test set which has 9 times higher OOV ratio, the perplexity reduction shows a similar trend. However, these reductions vanish when an in-domain test set is used. Note that we use the same treatment of OOV words for computing the perplexities which is used in KenLM (Heafield, 2013).

3.2 Analysis

Out-of-domain and Out-of-vocabulary We selected the Finnish language for which the number and ratio of OOVs are close on its out-of-domain and in-domain test sets (NT and EU), while showing substantial reduction in perplexity on out-of-domain test set, see FI bars on Figure 1. Figure 2 (left), shows the full perplexity results for Finnish for vanilla MKN, and our extensions when tested on in-domain (EU) and out-of-domain (NT) test sets. The discount plot, Figure 2 (middle) illustrates the behaviour of the various discount parameters. We also measured the average hit length for queries by varying m on in-domain and out-of-domain test sets. As illustrated in Figure 2 (right) the in-domain test set allows for longer matches to the training data as

[5] http://www.statmt.org/{wmt13,14,15}/test.tgz
[6] Streamed via Twitter API on 17/05/2016.
[7] http://www.statmt.org/wmt14/medical-task/

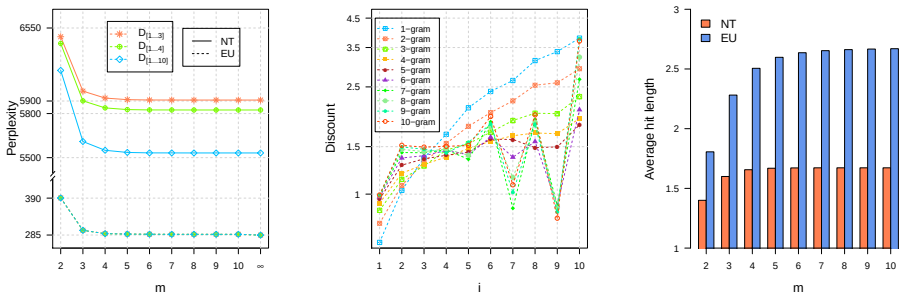

Figure 2: Statistics for the Finnish section of Europarl. The left plot illustrates the perplexity when tested on an out-of-domain (NT) and in-domain (EU) test sets varying LM order, m. The middle plot shows the discount parameters $D_{i\in[1...10]}$ for different m-gram orders. The right plot correspond to average hit length on EU and NT test sets.

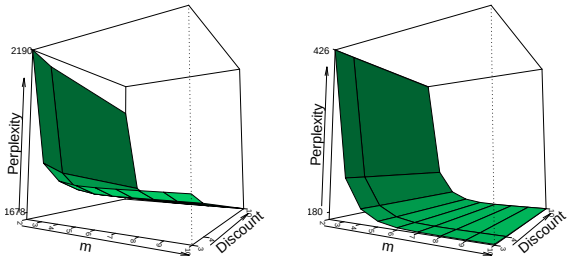

Figure 3: Perplexity (z-axis) vs. $m \in [2...10]$ (x-axis) vs. number of discounts $D_{i\in3,4,10}$ (y-axis) for German language trained on Europarl (left), and CommonCrawl2014 (right) and tested on news-test. Arrows show the direction of the increase.

m grows. This indicates that having more discount parameters is not only useful for test sets with extremely high number of OOV, but also allows for a more elegant way of assigning mass in the smoothing process when there is a domain mismatch.

Interdependency of m, data size, and discounts To explore the correlation between these factors we selected the German and investigated this correlation on two different training data sizes: Europarl (61M words), and CommonCrawl 2014 (984M words). Figure 3 illustrates the correlation between these factors using the same test set but with small and large training sets. Considering the slopes of the surfaces indicates that the small training data regime (left) which has higher sparsity, and more OOV in the test time benefits substantially from the more accurate discounting compared to the large training set (right) in which the gain from discounting is slight.[8]

[8]Nonetheless, the improvement in perplexity consistently grows with introducing more discount parameters even under the large training data regime, which suggests that more discount parameters, e.g., up to D_{30}, may be required for larger training corpus to reflect the fact that even an event with frequency of 30 might be considered rare in a corpus of nearly 1 billion words.

4 Conclusions

In this work we proposed a generalisation of Modified Kneser-Ney interpolative language modeling by introducing new discount parameters. We provide the mathematical proof for the discount bounds used in Modified Kneser-Ney and extend it further and illustrate the impact of our extension empirically on different Europarl languages using in-domain and out-of-domain test sets.

The empirical results on various training and test sets show that our proposed approach allows for a more elegant way of treating OOVs and mass assignments in interpolative smoothing. In future work, we will integrate our language model into the Moses machine translation pipeline to intrinsically measure its impact on translation qualities, which is of particular use for out-of-domain scenario.

Acknowledgements

This research was supported by the National ICT Australia (NICTA) and Australian Research Council Future Fellowship (project number FT130101105). This work was done when Ehsan Shareghi was an intern at IBM Research Australia.

A. Inequalities

We prove that these inequalities hold in expectation by making the reasonable assumption that events in

the natural language follow the power law (Clauset et al., 2009), $p\big(C(\mathbf{u}) = f\big) \propto f^{-1-\frac{1}{s_m}}$, where s_m is the parameter of the distribution, and $C(\mathbf{u})$ is the random variable denoting the frequency of the m-grams pattern $\mathbf{u}$. We now compute the expected number of unique patterns having a specific frequency $E[n_i[m]]$. Corresponding to each m-grams pattern $\mathbf{u}$, let us define a random variable $X_\mathbf{u}$ which is 1 if the frequency of $\mathbf{u}$ is i and zero otherwise. It is not hard to see that $n_i[m] = \sum_\mathbf{u} X_\mathbf{u}$, and

$$\begin{aligned}E\big[n_i[m]\big] &= E\big[\sum_\mathbf{u} X_\mathbf{u}\big] = \sum_\mathbf{u} E[X_\mathbf{u}] = \sigma^m E[X_\mathbf{u}] \\ &= \sigma^m\Big(p\big(C(\mathbf{u}) = i\big) \times 1 + p\big(C(\mathbf{u}) \neq i\big) \times 0\Big) \\ &\propto \sigma^m i^{-1-\frac{1}{s_m}}.\end{aligned}$$

We can verify that

$$\begin{aligned}(i+1)E\big[n_{i+1}[m]\big] < iE\big[n_i[m]\big] \Leftrightarrow \\ (i+1)\sigma^m(i+1)^{-1-\frac{1}{s_m}} < i\sigma^m i^{-1-\frac{1}{s_m}} \Leftrightarrow \\ i^{\frac{1}{s_m}} < (i+1)^{\frac{1}{s_m}}.\end{aligned}$$

which completes the proof of the inequalities.

B. Discount bounds proof sketch

The leave-one-out (leaving those m-grams which occurred only once) log-likelihood function of the interpolative smoothing is lower bounded by backoff model's (Ney et al., 1994), hence the estimated discounts for later can be considered as an approximation for the discounts of the former. Consider a backoff model with absolute discounting parameter D, were $P(w_i|w_{i-m+1}^{i-1})$ is defined as:

$$\begin{cases} \frac{c(w_{i-m+1}^i)-D}{c(w_{i-m+1}^{i-1})} & \text{if } c(w_{i-m+1}^i) > 0 \\ \frac{Dn_{1+}(w_{i-m+1}^{i-1}\,\bullet)}{c(w_{i-m+1}^{i-1})}\bar{P}(w^i|w_{i-m+2}^{i-1}) & \text{if } c(w_{i-m+1}^i) = 0 \end{cases}$$

where $n_{1+}(w_{i-m+1}^{i-1}\,\bullet)$ is the number of unique right contexts for the w_{i-m+1}^{i-1} pattern. Assume that for any choice of $0 < D < 1$ we can define $\bar{P}$ such that $P(w_i|w_{i_m+1}^{i-1})$ sums to 1. For readability we use the $\lambda(w_{i-m+1}^{i-1}) = \frac{n_{1+}(w_{i-m+1}^{i-1}\,\bullet)}{c(w_{i-m+1}^{i-1})-1}$ replacement. Following (Chen and Goodman, 1999), rewriting the leave-one-out log-likelihood for KN (Ney et al., 1994) to include more discounts (in this proof up to D_4), results in:

$$\begin{aligned}&\sum_{\substack{w_{i-m+1}^i \\ c(w_{i-m+1}^i)>4}} c(w_{i-m+1}^i)\log\frac{c(w_{i-m+1}^i)-1-D_4}{c(w_{i-m+1}^{i-1})-1} + \\ &\sum_{j=2}^{4}\Big(\sum_{\substack{w_{i-m+1}^i \\ c(w_{i-m+1}^i)=j}} c(w_{i-m+1}^i)\log\frac{c(w_{i-m+1}^i)-1-D_{j-1}}{c(w_{i-m+1}^{i-1})-1}\Big) + \\ &\sum_{\substack{w_{i-m+1}^i \\ c(w_{i-m+1}^i)=1}}\Big(c(w_{i-m+1}^i)\log(\sum_{j=1}^{4} n_j[m]D_j)\lambda(w_{i-m+1}^{i-1})\bar{P}\Big)\end{aligned}$$

which can be simplified to,

$$\begin{aligned}&\sum_{\substack{w_{i-m+1}^i \\ c(w_{i-m+1}^i)>4}} c(w_{i-m+1}^i)\log(c(w_{i-m+1}^i)-1-D_4) + \\ &\sum_{j=2}^{4}\Big(jn_j[m]\log(j-1-D_{j-1})\Big) + \\ &n_1[m]\log(\sum_{j=1}^{4} n_j[m]D_j) + \text{const}\end{aligned}$$

To find the optimal D_1, D_2, D_3, D_4 we set the partial derivatives to zero. For D_3,

$$\begin{aligned}&\frac{\partial}{\partial D_3} = n_1[m]\frac{n_3[m]}{\sum_{j=1}^4 n_j[m]D_j} - \frac{4n_4[m]}{3-D_3} = 0 \Rightarrow \\ &n_1[m]n_3[m](3-D_3) = 4n_4[m]\sum_{j=1}^{4} n_j[m]D_j \Rightarrow \\ &3n_1[m]n_3[m] - D_3 n_1[m]n_3[m] - 4n_4[m]n_1[m]D_1 > 0 \\ &\Rightarrow 3 - 4\frac{n_4[m]}{n_3[m]}D_1 > D_3 \quad \blacksquare\end{aligned}$$

And after taking $c(w_{i-m+1}^i) = 5$ out of the summation, for D_4:

$$\begin{aligned}&\frac{\partial}{\partial D_4} = \sum_{c(w_{i-m+1}^i)>5}\frac{-c(w_{i-m+1}^i)}{c(w_{i-m+1}^i)-1-D} - \frac{5n_5[m]}{4-D_4} \\ &+ n_1[m]\frac{n_4[m]}{\sum_{j=1}^4 n_j[m]D_j} = 0 \Rightarrow -\frac{5n_5[m]}{4-D_4} \\ &+ n_1[m]\frac{n_4[m]}{\sum_{j=1}^4 n_j[m]D_j} > 0 \Rightarrow n_1[m]n_4[m](4-D_4) \\ &> 5n_5[m]\sum_{j=1}^{4} n_j[m]D_j \Rightarrow 4 - 5\frac{n_5[m]}{n_4[m]}D_1 > D_4 \quad \blacksquare\end{aligned}$$

References

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Workshop on Statistical Machine Translation*.

Kenneth Heafield. 2013. *Efficient Language Modeling Algorithms with Applications to Statistical Machine Translation*. Ph.D. thesis, Carnegie Mellon University.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation summit*.

Hermann Ney, Ute Essen, and Reinhard Kneser. 1994. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech & Language*, 8(1):1–38.

Ehsan Shareghi, Matthias Petri, Gholamreza Haffari, and Trevor Cohn. 2015. Compact, efficient and unlimited capacity: Language modeling with compressed suffix trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at sixteen: Update and outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*.

Andreas Stolcke. 2002. SRILM–an extensible language modeling toolkit. In *Proceedings of the International Conference of Spoken Language Processing*.

Yee Whye Teh. 2006a. A Bayesian interpretation of interpolated Kneser-Ney. Technical report, NUS School of Computing.

Yee Whye Teh. 2006b. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.