BAYESIAN ESTIMATION METHODS FOR N-GRAM LANGUAGE MODEL ADAPTATION

Marcello Federico

IRST – Istituto per la Ricerca Scientica e Tecnologica
I–38050 Povo, Trento, Italy.

ABSTRACT

Stochastic n-gram language models have been successfully applied in continuous speech recognition for several years. Such language models provide many computational advantages but also require huge text corpora for parameter estimation. Moreover, the texts must exactly reflect, in a statistical sense, the user's language. Estimating a language model on a sample that is not representative severely affects speech recognition performance. A solution to this problem is provided by the Bayesian learning framework. Beyond the classical estimates, a Bayes derived interpolation model is proposed. Empirical comparisons have been carried out on a 10,000-word radiological reporting domain. Results are provided in terms of perplexity and recognition accuracy.

1. INTRODUCTION

Current continuous speech recognition systems combine an *acoustic model* and a *language model* with a Bayes' decision rule that minimizes the probability of error [2]. Given a sequence of acoustic observations O, e.g. feature vectors extracted by short-time spectral analysis of the speech signal, the corresponding word sequence W is computed according to the following decision rule:

$$\arg\max_{W} Pr(W)Pr(O \mid W). \tag{1}$$

Computation of formula (1) is performed by a search (or decoding) algorithm that exploits suitable decompositions of the involved probabilities. The acoustic model supplies the search algorithm with acoustic matching probabilities, i.e. the probability of any sub-sequence of O matching a given word or phoneme of the language. These probabilities are usually computed with Hidden Markov Models [4]. The language model instead provides linguistic probabilities (or scores) of words that constitute hypotheses of the search process. These probabilities are in general computed with n-gram language models (LMs). For instance, a 2-gram (bigram) LM approximates the probability of a word sequence $W = w_1, \ldots, w_m$ with the product:

$$Pr(w_1, \ldots, w_m) = Pr(w_1) \prod_{t=2}^{m} Pr(w_t \mid w_{t-1})$$

N-gram LMs have many advantages, mainly computational, but also some drawbacks: they need large text samples, to estimate parameters, that well represent the language of the user(s). In fact, moving from one application domain to another induces changes in the lexicon and in the statistical features of the language. In this paper, it will be shown that even inside a very specific application domain (radiological reporting) two user groups (hospitals) may employ statistically different languages.

Estimating a language model on a text sample that is not representative can severely affect recognition performance. This problem can be tackled with adaptation techniques. Hence, a language model can be first estimated on a very large user-independent (UI) text corpus and is then gradually adapted as user-dependent (UD) linguistic data become available.

In this work, different Bayesian learning [2] approaches to bigram language model adaptation are discussed and compared.

2. LANGUAGE MODEL ADAPTATION

Language model adaptation can be seen as an estimation problem in which a parametric model must be determined, given a UD adaptation sample S, usually very small, and some *a priori* knowledge. Typically, such knowledge is extracted from a large UI text sample S'.

As the objective is to estimate $Pr(z \mid y)$ given a subsample of S, S_y, containing all bigrams beginning with y, the context variable y will be omitted in the following.

2.1. Bayes and MAP Estimates

From $Pr(z)$ belonging to the family of discrete distributions $Pr(z; \theta_1, \ldots, \theta_r)$ on the population $V = \{1, \ldots, r\}$, point estimates can be derived from the posterior distribution either with the Bayesian criterion:

$$\theta^B = E[\theta \mid S]$$

or with the maximum *a posteriori* (MAP) criterion:

$$\theta^M = arg \max_{\theta} Pr(\theta \mid S).$$

From basic statistical methods [2], the following formulas can be derived by considering a Dirichlet prior distribution estimated on the UI sample S':

$$Pr^M(z) = \theta_z^M = (\frac{m}{m + \epsilon m'}) f(z) + (\frac{\epsilon m'}{m + \epsilon m'}) f'(z) \quad (2)$$

$$Pr^B(z) = \theta_z^B = (\frac{m}{m + \epsilon(m' + r)}) f(z) + (\frac{\epsilon(m' + r)}{m + \epsilon(m' + r)}) \hat{f}'(z). \quad (3)$$

where m and m' are the sizes of S and S', f and f' are the relative frequencies of S and S', $\hat{f}'$ is the smoothed relative frequency of S' obtained by adding 1 to all frequencies, and ϵ is a constant such that $0 \leq \epsilon \leq 1$.

Both estimates result in a combination of frequencies of the two samples S and S', with weights proportional to the sample sizes. When a large adaptation sample is used (i.e. $m \to \infty$), both formulas asymptotically converge to the maximum likelihood estimate of θ_v (i.e. $f(v)$).

The ϵ constant has been introduced to reduce the bias of the estimates. In fact, the Bayesian (B) and the MAP (M) estimates both require $\epsilon = 1$ but, as in LM adaptation it usually is $m' >> m$, the influence of the UI sample S' might overwhelm that of the UD sample S. Hence, by setting ϵ in the interval $[0, 1]$ the relative weight of the adaptation sample can be increased.

2.2. Bayes and MAP Interpolation

According to formulas (2) and (3), the following MAP Interpolation (MI) and Bayes Interpolation (BI) parametric models can be derived:

$$Pr^{MI}(z) = \theta_z^{MI} = \lambda f(z) + (1 - \lambda) f'(z) \quad (4)$$

$$Pr^{BI}(z) = \theta_z^{BI} = \lambda f(z) + (1 - \lambda) \hat{f}'(z) \quad (5)$$

where the weight $0 \leq \lambda \leq 1$ has to be determined. It is known that the weight of the above mixtures can be estimated from a sample S by means of the EM (*expectation maximization*) algorithm [1], which numerically approximates their maximum likelihood estimates.

3. EMPIRICAL COMPARISON

3.1. Application domain

Experiments have been performed on two text samples containing radiological reports produced by two hospitals. Sample A contains about 1.9 million word occurrences while sample B about 600,000. The vocabulary sizes are, respectively, of 10,370 and of 6,487 words, for a total of 11,170 different words. Besides the lexical differences, the two samples provide a very different distribution of report types. In fact, sample A is very much focused on *x-ray* reports, and particularly on *x-ray chest* reports, while sample B covers the possible types of reports in a more balanced way [1]. The different distributions of report types indeed affect the n-gram frequency distributions. For instance, in Figure 1 a plot of the relative frequencies of all words (1-grams) observed in both corpora is shown. The shape of the cloud around the diagonal axis shows that larger (relative) differences among frequencies can be observed especially for the low frequent words, which are in fact the majority.

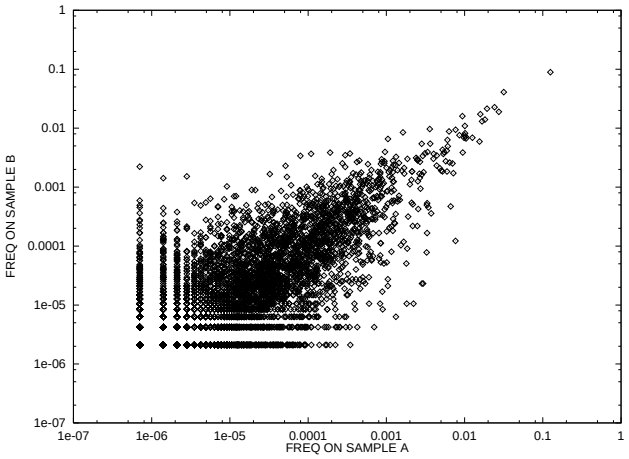

Figure 1: Unigram relative frequencies measured on the two samples A and B. Relative frequencies are plotted for all words appearing in both samples. The column and row shaped plots, close to the origin, correspond to words occurring once, twice, etc. in sample B and A, respectively.

In the following, LM adaptation experiments will be presented where sample A provides the UI sample data (i.e. S') and sample B the UD adaptation data (i.e. S). In order to carry out comparison tests, the sample B is before split into training and testing subsamples in the proportion of 3:1 (roughly 450,000 and 150,000 words). In the experiments, adaptation algorithms have been applied with increasing sizes of adaptation data, with corresponding performances measured on the testing data.

3.2. Compared LMs

Adaptation methods have been compared by employing the following linearly smoothed bigram LMs:

B: $Pr(z \mid y) = \lambda_1([y]) Pr^B(z \mid y) + \lambda_2([y]) Pr^{BI}(z)$

M: $Pr(z \mid y) = \lambda_1([y]) Pr^M(z \mid y) + \lambda_2([y]) Pr^{BI}(z)$

MI: $Pr(z \mid y) = \lambda_1([y]) Pr^{MI}(z \mid y) + \lambda_2([y]) Pr^{BI}(z)$

satisfying:

$$\forall y \in V : \lambda_i([y]) \geq 0 \; i = 1, 2, \sum_{i=1,2} \lambda_i([y]) = 1. \quad (6)$$

[1] Sample A is related to emergency examinations while sample B to general examinations.

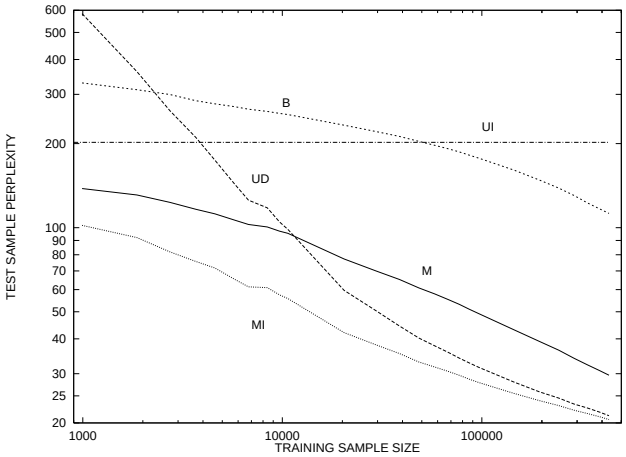

Figure 2: Learning curves of the MAP (M), Bayesian (B), and MAP Interpolation (MI) adaptation methods in the bigram case. The UD curve refers to the LM estimated on the adaptation sample S only; the UI line refers to the bigram LM estimated on the user-independent sample S' only.

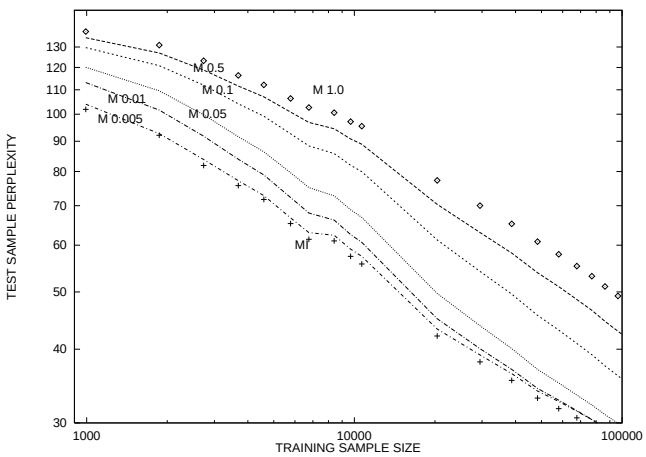

Figure 3: MAP adaptation method in the bigram case with different values of ϵ. The curves of the pure MAP (i.e. M with $\epsilon = 1$) and MI adaptation methods are also plotted. The MI adaptation curve results to be a lower bound for the MAP method.

The Bayesian (B), MAP (M) and MAP Interpolation (MI) estimates in the above LMs are the same defined in Section 2 with the samples S_y and S'_y replacing S and S'. Preliminary experiments showed that Bayes Interpolation (BI) is very suited to unigram adaptation but does not significantly differ from the standard B method in the bigram case. For this reason BI is not included in the tested bigram LMs. Finally, for the sake of comparison, the UD and UI bigram LMs representing, respectively, the lack of *a priori* information and of adaptation data were also considered:

$$\begin{aligned} \text{UD:} \quad & Pr(z \mid y) = \lambda_1([y]) f(z \mid y) + \lambda_2([y]) \hat{f}(z) \\ \text{UI:} \quad & Pr(z \mid y) = \lambda_1([y]) f'(z \mid y) + \lambda_2([y]) \hat{f}'(z) \\ & \text{satisfying (6).} \end{aligned}$$

As in adaptation training data are generally few, the number of interpolation parameters is controlled by considering equivalence classes of V, depending on the absolute frequency $c(\cdot)$ of each word y in S:

$$\forall y \in V \quad [y] = \begin{cases} 0 & \text{if } c(y) \leq k_1 \\ c(y) & \text{if } k_1 < c(y) \leq k_2 \\ y + k_2 & \text{if } k_2 < c(y) \end{cases} \qquad (7)$$

The values of the thresholds k_1 and k_2 were empirically set to 2 and 10, respectively. Extensive experiments have shown that values for these thresholds are not critical and that good performance can be achieved with more than one assignment. In this way, frequent contexts have individual parameters, while all the other share parameters with other not frequent contexts. It is also easy to see that the number of equivalent classes in (7) varies with the size of S.
Besides, by eliminating redundant parameters the MI LM has been rewritten as:

$$\begin{aligned} \text{MI:} \quad Pr(z \mid y) = \; & \lambda_1([y]) f(z \mid y) + \lambda_2([y]) f'(z \mid y) + \\ & + \lambda_3([y]) Pr^I(z) \end{aligned}$$

satisfying (6) extended to $i = 1, 2, 3$.

Estimation of the interpolation parameters of all the above bigram LMs has been performed with the *Stacked Estimation* algorithm [3], which is a cross-validation based version of the EM method.

3.3. Perplexity Tests

Comparisons of the adaptive LMs in terms of perplexity provided the learning curves shown in Figure 2. The MI method appears to be superior than its competitors. In particular, B adaptation is particularly worse than M adaptation due to the effect of smoothing, which sensibly alters the *a priori* relative frequencies. In general, the advantage of adaptation is evident. For training size less than 10,000 words MI adaptation significantly improve over the UD LM and still performs better when all the adaptation data are used.
Further experiments were performed with M adaptation as a function of ϵ. Some of the computed learning curves are shown in Figure 3. As expected, improvements are achieved by reducing the value of ϵ. In fact, convergence was found around the value 0.005, where the M LM performs almost like the MI one. Further reductions of ϵ, e.g. to 0.001, started to worsen performance and are not included in the plot. Interestingly, this shows that the MI curve is like a sort of lower bound for the M method. An explanation for this is that the bigram scheme employing the M method, has, thanks to the ϵ parameter, enough degrees of freedom to behave like the MI scheme (8). In favor of the MI method is, of course, the fact that all the parameters are estimated automatically.

3.4. Speech Recognition Tests

Speech recognition experiments on bigram LM adaptation have been carried out with the speaker-independent real-time

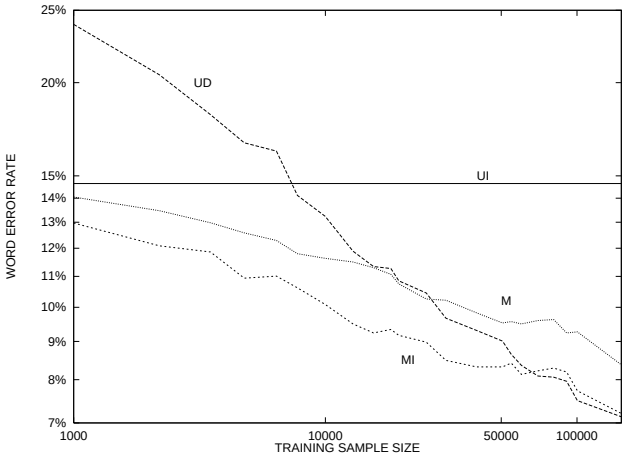

Figure 4: Continuous speech recognition tests comparing the Interpolation and the MAP adaptation methods.

continuous-speech recognizer developed at IRST [3]. The task considered consists of 150 reports dictated by two male speakers. The reports were selected inside the test subsample of the UD corpus. Performance of each LM has been evaluated with the word error rate (WER) produced by the recognizer employing the LM. All the tested bigram LMs have been represented in terms of a probabilistic finite state network [3].

The performed tests aimed at evaluating LM adaptation especially in the short term, that is with UD adaptation examples of few thousands of words. In fact, the main purpose of adaptation is to improve performance of an "out-of-the-box" recognizer, that has only been trained on some general UI text corpus. During usage of the system new texts are produced which reflect the user language. These texts are exploited to adapt the LM of the recognizer and therefore to improve performance of the system.

The best two performing adaptation methods were compared: the MI and the M ($\epsilon = 1$) one. Again, as references the UI and UD LMs were tested. The resulting learning curves are shown in Figure 4, while the corresponding WERs are shown in Table 1.

As expected, the speech recognition performance evaluation maintains the same rank orders provided by the perplexity tests, but with significantly lower relative distances. Again, the LM providing the best learning curve is the one adapted with Interpolation. Convergence with the UD LM occurs when around 50,000 words of training data are available. After that point the two LMs seem to behave similarly. To verify this hypothesis, for every training data size, all pairwise WER differences between the considered LMs were statistically verified with a means test for paired samples [5].

By comparing the UI LM estimated on the adaptation sample only and the one adapted with the MI method, the WER is reduced by 11% (i.e. from 14.65% to 12.98%) after only 1,000 words of training data, and by more than 25% after 5,000 words. With respect to the UD LM only using user-independent training material, WER relative differences ranging from 24% to 46% are observed if less than 10,000 training words are used.

Train size	*UD*	*MI*	*M*
1,000	23.98	12.98	14.06
2,000	20.48	12.09	13.47
3,500	18.12	11.86	12.98
4,700	16.61	10.94	12.58
6,400	16.19	10.62	11.80
10,000	13.24	10.09	11.63
20,000	$_1$ 10.85	9.17	$_1$ 10.75
30,000	$_1$ 9.67	8.49	$_1$ 10.22
40,000	$_1$ 9.31	8.32	$_1$ 9.83
50,000	$_{1,2}$ 9.01	$_1$ 8.32	$_2$ 9.53
100,000	$_1$ 7.50	$_1$ 7.73	9.27
200,000	$_{1,2}$ 6.88	$_1$ 6.85	$_2$ 7.78
300,000	$_{1,2}$ 6.88	$_1$ 6.29	$_2$ 7.50
400,000	$_{1,2}$ 6.52	$_1$ 6.16	$_2$ 7.24
500,000	$_{1,2}$ 6.29	$_1$ 6.03	$_2$ 7.11

Table 1: WER measured with different LMs by employing increasing UD adaptation data. Subscripts in each row indicate WER pairs which do not significatively differ.

4. CONCLUSIONS

The conclusion is that LM adaptation techniques can be successfully employed to improve performance of a speech recognizer when little training data of the user are available. In particular, MAP Interpolation adaptation outperforms the classical Bayes' derived estimates. Moreover, MAP Interpolation adaptation is conservative, in the sense that if training data become large, performance does not deteriorate with respect to maximum likelihood estimation.

5. REFERENCES

1. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38, 1977.
2. R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, NY, 1973.
3. M. Federico, M. Cettolo, F. Brugnara, and G. Antoniol. Language modeling for efficient beam-search. *Computer Speech and Language*, 9:353–379, 1995.
4. F. Jelinek and R. L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Pattern Recognition in Practice*, pages 381–397, Amsterdam, Holland, 1980.
5. R. A. Johnson and D. W. Wichern, eds. *Applied Multivariate Statistical Analysis*. Prentice Hall, Englewood Cliffs, NJ, 1992.