EFFICIENT LANGUAGE MODEL ADAPTATION THROUGH MDI ESTIMATION

Marcello Federico

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
I-38050 Povo, Trento, Italy.
federico@itc.it

ABSTRACT

This paper presents a method for n-gram language model adaptation based on the principle of minimum discrimination information. A background language model is adapted to fit constraints on its marginal distributions that are derived from new observed data. This work gives a different derivation of the model by Kneser *et al.* (1997) and extends its application to interpolated language models. The proposed method has been evaluated on an Italian 60K-word broadcast news task.

Keywords: language model adaptation, minimum discrimination information estimation, generalized iterative scaling.

1. INTRODUCTION

Current speech recognition systems model the recognition process as a statistical decision criterion. Two main information sources are usually combined: an acoustic model and a language model (LM). Criteria used for training the LM - e.g. maximum likelihood - generally provide parameter estimates that perform well on the average, assuming that testing and training data are generated from similar sources. However, the resulting parameters are in general not able to cope with significant variations of the language, such as domain changes, topic shifts, and user changes.

This work presents a technique for language model adaptation based on the statistical principle of minimum discrimination information (MDI). Briefly, a new LM is estimated so that it is "as close as possible" to a general background LM and satisfies some constraints empirically derived from a relatively small adaptation sample. In fact, given a consistent set of constraints, a solution can be computed that minimizes the Kullback-Leibler distance from the background joint distribution, and satisfies the constraints.

Previous work on MDI adaptation or affine techniques has been presented in [**?**, **?**, **?**, **?**, **?**]. In general, the literature reported improvements of performance but at very high computational costs. Recently, [**?**] presented a simple and fast MDI based technique to adapt a backing-off n-gram LM [**?**], by using constraints on marginal distributions. The present work takes inspiration from [**?**], gives a different derivation of its result, and extends it to interpolated LMs. Experimental results are reported for a 60K-word Italian broadcast news recognition task.

2. BACKGROUND LM

An n-gram LM approximates the probability $\Pr(W_1^T)$ of a text of words $W_1^T = w_1 \ldots, w_t, \ldots, w_T$, from a finite vocabulary V, with the product:

$$\Pr(W_1^T) = \prod_{t=1}^{T} \Pr(w_t \mid h_t) \qquad (1)$$

where $h_t = w_{t-n+1} \ldots w_{t-1}$. Hence, the history h_t limits the dependence of word w_t to the $n-1$ words preceding it. Data sparseness of real texts suggest to *smooth* n-gram probabilities. Given a training text corpus B, the conditional probability of an n-gram $hw \in V^n$ can be expressed by *interpolation* as follows:

$$P_B(w \mid h) = f_B^*(w \mid h) + \lambda_B(h) P_B(w \mid h') \qquad (2)$$

where $f_B^*(w \mid h)$ is the *discounted* frequency, $\lambda_B(h)$ is the zero-frequency probability:

$$\lambda_B(h) = 1.0 \; - \sum_{w \in V} f_B^*(w \mid h),$$

and h' is the lower order history, obtained by shortening h by one. One discounting method is that introduced by [**?**]:

$$f_B^*(w \mid h) = max \left\{ \frac{c_B(hw) - \beta}{c_B(h)}, 0 \right\} \qquad (3)$$

with

$$\beta = \frac{n_1}{n_1 + 2n_2}$$

where $c_B(\cdot)$ is the counting function on sample B, and n_i represents the number of n-grams that occurred exactly i times in B.

3. MDI LM ADAPTATION

The basic approach of MDI adaptation is shown in Figure **??**. A background LM is estimated on a large corpus B. As adaptation texts A become available, features are extracted from A to constrain the adapted model. Formally, a set of linear constraints on the joint distribution $P_A(h, w)$ is specified, i.e.:

$$\sum_{hw \in V^n} P_A(h, w) \delta_i(hw) = \hat{P}_A(S_i) \qquad i = 1, \ldots, M \qquad (4)$$

where $\delta_i(\cdot)$ are indicator functions of subsets $S_i \subset V^n$, also called features, and $\hat{P}_A(S_i)$ are empirical estimates of the features on A. Finally, the joint distribution $P_A(\cdot)$ is defined as the distribution that satisfies the constraints (**??**) and minimizes the Kullback-Leibler distance [**?**] from the background joint distribution, i.e.:

$$P_A(\cdot) = \arg\min_{Q(\cdot)} \sum_{hw \in V^n} Q(h,w) \log \frac{Q(h,w)}{P_B(h,w)}. \quad (5)$$

Assuming the constraints (**??**) are consistent and that there is an integer k so that each $hw \in V^n$ satisfies exactly k features, it can be shown [**?**] that an iterative algorithm exists that converges to a unique solution of (**??**). The algorithm, called Generalized Iterative Scaling (GIS) [**?**], performs the following iterations:

$$P_A^{(0)}(h,w) = P_B(h,w) \quad (6)$$

$$P_A^{(r+1)}(h,w) = P_A^{(r)}(h,w) \prod_{i=1}^{M} \left(\frac{\hat{P}_A(S_i)}{P_A^{(r)}(S_i)} \right)^{\frac{\delta_i(hw)}{k}} \quad (7)$$

where:

$$P_A^{(r)}(S_i) = \sum_{hw \in V^n} P_A^{(r)}(h,w)\delta_i(hw) \qquad i = 1, \ldots, M. \quad (8)$$

Given that the adaptation sample is typically small, one may assume than only unigram features can be reliably estimated on A. Hence, the following constraints can be set:

$$\sum_{hw \in V^n} P_A(h,w)\delta_{\hat{w}}(hw) = \hat{P}_A(\hat{w}) \qquad \forall \hat{w} \in V \quad (9)$$

where $\delta_{\hat{w}}(hw) = 1$ if $w = \hat{w}$ and 0 otherwise, and $\hat{P}_A(\cdot)$ is a unigram LM on A. It can be shown that the constraints (**??**) are consistent and define exactly one partition of V^n, i.e. $k = 1$. The GIS algorithm reduces to the following closed form:

$$P_A(h,w) = P_B(h,w)\ \alpha(w) \quad (10)$$

where:

$$\alpha(w) = \frac{\hat{P}_A(w)}{P_B(w)}. \quad (11)$$

The objective of LM adaptation is finally to estimate the conditional distribution $P_A(w \mid h)$, given the adaptation sample A and the conditional background LM $P_B(w \mid h)$. By elementary probability theory, the conditional version of (**??**) can be derived:

$$P_A(w \mid h) = \frac{P_B(w \mid h)\ P_B(h)\ \alpha(w)}{\sum_{\hat{w} \in V} P_B(\hat{w} \mid h)\ P_B(h)\ \alpha(\hat{w})} \quad (12)$$

$$= \frac{P_B(w \mid h)\ \alpha(w)}{\sum_{\hat{w} \in V} P_B(\hat{w} \mid h)\ \alpha(\hat{w})}. \quad (13)$$

Improvements on the adaptation model can be obtained by exponentially smoothing the scaling factor $\alpha(\cdot)$ [**?**]. Hence, a parameter γ is introduced in the formula (**??**), i.e.:

$$\alpha(w) = \left(\frac{\hat{P}_A(w)}{P_B(w)} \right)^{\gamma} \quad 0 < \gamma \leq 1 \quad (14)$$

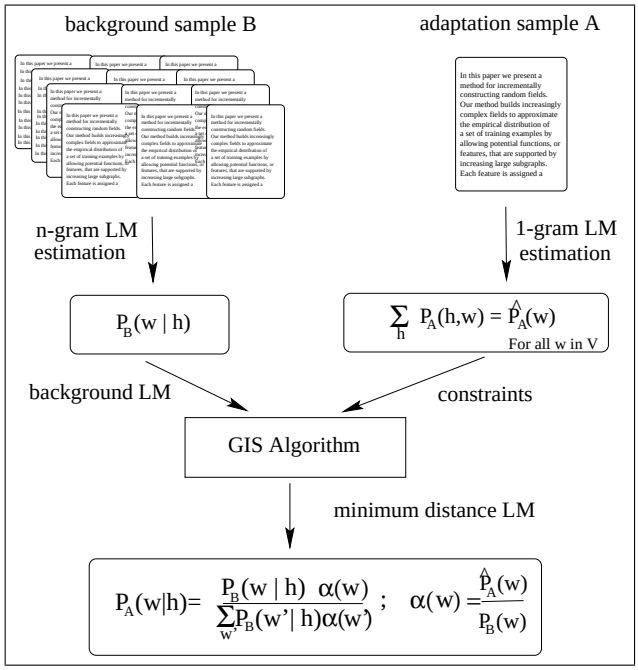

Figure 1: LM adaptation through MDI.

that can be empirically estimated. Intuitively, values of γ less than one reduce the effect of the adaptation distribution $\hat{P}_A(\cdot)$ in the ratio (**??**). This is analogue to what occurs in the Bayesian adaptation framework [**?**], where a constant factor is introduced to reduce the bias of the estimator.

4. IMPLEMENTATION

At first glance, the adapted n-gram LM (**??**) requires computing a normalization term by $|V|$ operations. A speed-up in the computation of the normalization can be obtained by exploiting the structure of an interpolated language model. Let one assume that the background LM is represented by equation (**??**).

Hence, the adapted LM (**??**) can be expressed by an interpolation as well:

$$P_A(w \mid h) = f_A^*(w \mid h) + \lambda_A(h) P_A(w \mid h') \quad (15)$$

where:

$$f_A^*(w \mid h) = \frac{f_B^*(w \mid h)\alpha(w)}{z(h)}$$

$$\lambda_A(h) = \frac{\lambda_B(h) z(h')}{z(h)}$$

$$z(h) = \sum_{w: f_B(h,w) > 0} f_B^*(w \mid h)\alpha(w) + \lambda_B(h) z(h')$$

with the following initial values for the empty history ϵ:

$$z(\epsilon) = \sum_w P_B(w)\alpha(w)$$

$$P_A(w \mid \epsilon) = P_B(w)\alpha(w) z(\epsilon)^{-1}$$

In (**??**), the normalization is carried out recursively by the term $z(\cdot)$. Given an history h, of length $n-1$, the recursion step requires summing over the set of n-grams hw that

#	*period*	*# recordings*	*# texts*
1	28 Sept - 2 Oct	14 (4':12")	63 (3,086 w.)
2	4 Oct - 9 Oct	21 (4':53")	50 (2,263 w.)
3	12 Oct - 16 Oct	25 (6':38")	40 (2,036 w.)
4	19 Oct - 23 Oct	13 (4':20")	29 (1,489 w.)
5	26 Oct - 30 Oct	0	35 (1,311 w.)
6	2 Nov - 6 Nov	10 (2':28")	47 (1,809 w.)
7	9 Nov - 13 Nov	18 (3':52")	33 (1,509 w.)
8	16 Nov - 20 Nov	18 (5':50")	36 (2,257 w.)
9	23 Nov - 27 Nov	18 (4':57")	39 (1,602 w.)
10	30 Nov - 4 Dec	15 (3':25")	33 (1,771 w.)
Totals		152 (40':34")	405 (19,133 w.)

Table 1: Statistics of the broadcast news recordings and texts used for the experiments.

were *observed* in sample B. Practically speaking, for large vocabulary applications the size of this set decreases with n and is, on the average, much smaller than $|V|$. The recursion ends with the term $z(\epsilon)$, whose computation, requiring $|V|$ steps, can be performed off-line only once.

Statistics about the number of computations required by the normalization term were collected on a 34M-word corpus with a 60K-word vocabulary. It results that a bigram LM requires 90% of the time less than 1000 sums, while a trigram LM requires the same amount plus up to 10 sums for normalizing the trigram probability.

5. EXPERIMENTAL RESULTS

LM adaptation experiments have been conducted on a radio broadcast news task. The testing set consists of about 30 recordings of the Italian GR1 news program, that were manually segmented into 152 single news. The recordings span over the period October-December 1998. Moreover, a larger text corpus of the same news program was collected through the Internet. As background LM, a 60K-word bigram LM was estimated on a 34M-word corpus containing 1997 issues of the Italian financial newspaper *Il Sole 24 Ore*. The out-of-vocabulary word rate of the LM is, wrt to the broadcast news texts, about 2%.

Experiments were conducted by employing the IRST speech recognizer trained in a multi-speaker modality on the same recordings. Overlap of training and testing material was avoided by employing a cross-validation scheme. Background and adapted LMs have been represented by recursive probabilistic networks [**?**] that exploit pre-compiled grammars modeling expressions like numbers, dates, percentages, etc. The speech recognition experiments run on a Pentium II 400MHz in twice the real-time and took 200Mb of process size.

Supervised LM adaptation has been performed incrementally with steps of one week. Hence, the available audio recordings and texts were grouped into 10 blocks of one week. Details about each data-block are shown in Table **??**. (Note that no recordings are available for week 5.) The background LM is adapted to model week W_n by employing all the past available texts $W_1, W_2, ..., W_{n-1}$ as adaptation sample.

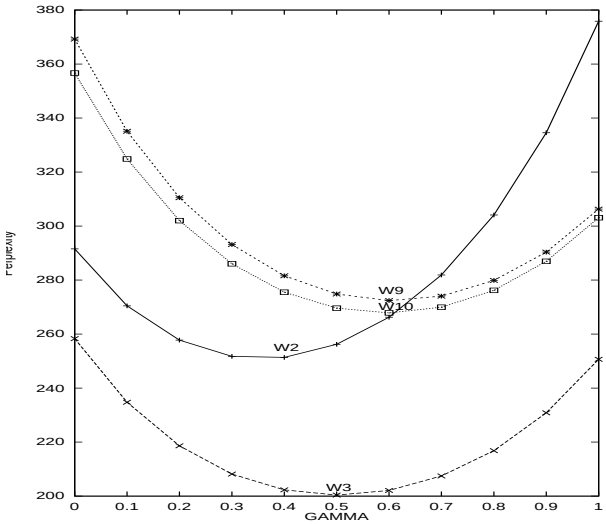

Figure 2: Perplexity for different values of the parameter γ. $\gamma = 0$ is equivalent to non adaptation, $\gamma = 1$ corresponds to the standard GIS solution.

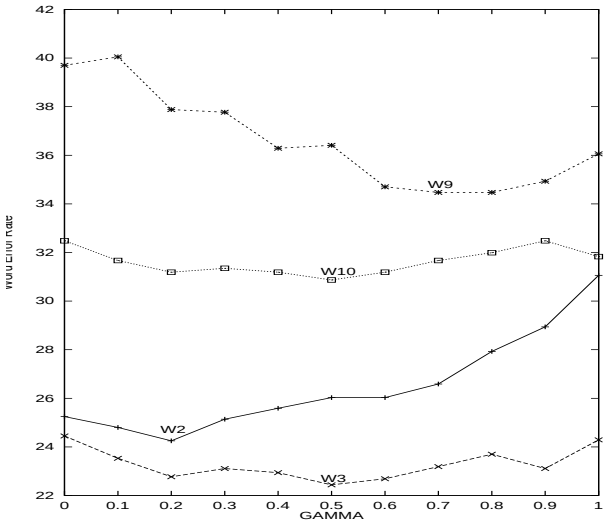

Figure 3: Word error rate for different values of γ.

5.1. Fixed vocabulary

In order to compare adaptation performance by keeping the number of LM parameters fixed, the vocabulary was not augmented by new words observed in the adaptation data. Hence, perplexity based evaluation was conducted by letting the tuning parameter γ vary in the interval $[0, 1]$ with step 0.1. Performance curves of four data blocks are shown in Figure **??**. In general, all the test blocks provided similar trends. Significant perplexity reductions were observed, ranging from 12% to 25%. Larger improvements occur on higher week blocks, for which more adaptation material is available. Moreover, a slight upward shift of the optimal value of the parameter γ is evident, which probably means that the scaling factor $\alpha(\cdot)$ becomes more reliable.

Performance in terms of word error rate (WER) on the same four data blocks are shown in Figure **??**. It can

	$\gamma = 0$		$\gamma = .5$		$\gamma = .5 + \Delta V$	
W	PP	WER	PP	WER	ΔOOV	WER
1	320	25.8%	320	25.8%	-0.0%	25.8%
2	291	25.2%	256	26.0%	-0.1%	25.8%
3	258	24.4%	200	22.4%	-0.4%	22.3%
4	300	27.9%	229	26.3%	-0.2%	26.1%
5	335	-	262	-	-0.3%	-
6	270	26.9%	225	26.2%	-0.4%	26.0%
7	250	25.1%	203	23.7%	-0.6%	23.9%
8	266	30.5%	201	27.8%	-1.3%	27.8%
9	369	39.7%	275	36.4%	-1.2%	36.0%
10	357	32.5%	270	30.9%	-1.9%	28.5%
AV	298	28.6%	245	27.1%	-0.7%	26.9%

Table 2: Performance on all data sets with no LM adaptation ($\gamma = 0$), with adaptation ($\gamma = 0.5$), and with adaptation plus vocabulary augmentation.

be noted that the WER shows less regular curves and more limited variations. In fact, it is well known from the literature that the relationship between perplexity and WER is not obvious, and that there is some evidence of a quadratic relationship between their variations. However, the results on all the tested data blocks show that optimal values of γ are the same or close to those obtained with the perplexity. Testing blocks which exploit more adaptation data show more stable trends and larger WER reductions.

By assuming $\gamma = 0.5$ as a good guess on-the-average, perplexity and WER performance of the non adapted and adapted LMs are reported in the first two columns of Table **??**. Starting from week 7, significant improvements are evident both in terms of perplexity and WER. Relative variations range between 5% and 8.8%. On the average, the global WER reduction is about 5%.

5.2. Augmented vocabulary

It can be of interest the contribution that new words found within the adaptation data can give in terms of recognition performance. It is known that out-of-vocabulary (OOV) words severely affect recognition accuracy. Empirically observations tell that one OOV word causes on the average about 1.5 recognition errors. Experiments with the vocabulary augmented with new words found within the adaptation data were carried out and results are shown in the last two columns of Table **??**. In particular, for each testing block the absolute OOV word rate variation and the WER are shown after augmenting the vocabulary and adapting the LM. As expected, augmented vocabulary adaptation increases performance on the average. In fact, global performance is improved by a further 0.8%. However, the average absolute variation of the WER (-0.2%) is almost four times lower than the corresponding variation of the OOV rate (-0.7%). As the inserted words are in general rare, it means that such words have about 25% of chance to be correctly recognized.

6. CONCLUSION

In this paper an efficient method was presented for adapting an n-gram LM to a new domain, topic, or speaker. The adapted model is obtained by taking a general background LM and by applying a non-linear scaling factor to its conditional distribution. The scaling factor is estimated so that the resulting model satisfies constraints on the marginal unigram distribution, that are derived from some relatively small data set representing the speaker's language. The key issues are the efficient way of rescaling and normalizing the conditional distribution, which exploits the structure of the interpolation based LM, and a smoothing parameter that allows to tune the effect of the scaling factor.

7. REFERENCES

[1] F. Brugnara and M. Federico. Dynamic language models for interactive speech applications. In *Proc. of EUROSPEECH*, pages 2751–2754, Rhodes, Greece, 1997.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, 1991.

[3] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

[4] S. A. Della Pietra, V. J. Della Pietra, R. Mercer, and S. Roukos. Adaptive language model estimation using minimum discrimination estimation. In *Proc. of ICASSP*, volume I, pages 633–636, San Francisco, CA, 1992.

[5] M. Federico. Bayesian estimation methods of n-gram language model adaptation. In *Proc. of ICSLP*, pages 240–243, Philadelphia, PA, 1996.

[6] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. Acoust., Speech and Signal Proc.*, ASSP-35(3):400–401, 1987.

[7] R. Kneser, J. Peters, and D. Klakow. Language model adaptation using dynamic marginals. In *Proc. of EUROSPEECH*, pages 1971–1974, Rhodes, Greece, 1997.

[8] R. Lau, R. Rosenfeld, and S. Roukos. Trigger-based language models: a maximum entropy approach. In *Proc. of ICASSP*, volume 2, pages 45–48, Minneapolis, MN, 1993.

[9] H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8:1–38, 1994.

[10] P. S. Rao, S. Dharanipragada, and S. Roukos. MDI adaptation of language models across corpora. In *Proc. of EUROSPEECH*, pages 1979–1982, Rhodes, Greece, 1997.

[11] P. S. Rao, M. D. Monkowski, and S. Roukos. Language adaptation via minimum discrimination information. In *Proc. of ICASSP*, volume 1, pages 161–164, Detroit, MI, 1995.

[12] R. Rosenfeld. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. PhD thesis, School of Computer Science - Carnegie Mellon University, Pittsburgh, PA, 1994.