Optimizing the Integration of a Statistical Language Model in HMM based Offline Handwritten Text Recognition

Matthias Zimmermann and Horst Bunke
Department of Computer Science, University of Bern
Neubrückstrasse 10, CH-3012 Bern, Switzerland
{zimmerma,bunke}@iam.unibe.ch

Abstract

Although handwritten text recognition has been studied for some years, only few authors have used statistical language models to increase the performance of their recognizers. In those few cases where a language model has been used, its integration has not been systematically optimized. In this paper we investigate the optimization of the integration of statistical language models into HMM based recognition systems for offline handwritten text. Based on experiments with the IAM database we show that the recognition performance of a general offline handwritten text recognizer can be substantially improved.

1 Introduction

Offline recognition of handwritten text has been investigated since many years [2, 4, 11]. But only recently authors have started to take advantage of the integration of statistical language models [5, 8, 12]. All these works are based on the HMM framework which directly supports the integration of such language models.

The goal of handwritten text recognition is to find the most likely sentence $\hat{W} = (w_1, w_2, \ldots, w_n)$ for a given observation sequence $X = (X_1, X_2, \ldots, X_m)$ provided by some feature extraction mechanism.

$$\hat{W} = \underset{W}{argmax}\ p(W|X) \tag{1}$$

Since HMM based classifiers compute an estimate of $p(X|W)$ for a given hypothesis W the Bayes' rule can be applied to rewrite Eq. (1) as follows.

$$\hat{W} = \underset{W}{argmax}\ p(X|W)p(W) \tag{2}$$

Consequently, the result of the HMM classification needs to be combined with the sentence probability $p(W)$. Eq. (2) can therefore be seen as the decomposition of Eq. (1) into the "optical" model $p(X|W)$ and a statistical language model represented by $p(W)$.

For the statistical language model most often so called n-gram models are used in the domain of speech recognition [10]. In the area of handwriting recognition, bigram models ($n = 2$) were used by [5, 8], and [12] compared the impact of both bigram and trigram ($n = 3$) language models on the recognizers' performance. Given a bigram language model, the sentence probability $p(W)$ is computed as the product $\prod p(w_i|w_{i-1})$, where $p(w_i|w_{i-1})$ stands for the conditional probability of word w_i directly following word w_{i-1}[1]. This decomposition allows for a simple and efficient integration of such language models into the Viterbi decoding step:

$$\phi_i = \phi_{i-1} + \log p(X_i|w_i) + \log p(w_i|w_{i-1}) \tag{3}$$

where X_i represents the observation sequence associated with word w_i, $\phi_0 = 0$ and $\phi_n = p(W|X)$ according to Eq. (1).

Since both the HMM and the n-gram language model only produce approximations of probabilities, we use two additional parameters, α and β. Their aim is to partially compensate the deficiencies of the optical model and the language model. This leads to

$$\phi_i = \phi_{i-1} + \log p(X_i|w_i) + \alpha\ \log p(w_i|w_{i-1}) + \beta \tag{4}$$

In this paper we will use the term *Grammar Scale Factor*[2] (GSF) for parameter α and *Word Insertion Penalty* (WIP) for parameter β. The GSF can be used to weight the influence of the language model against the optical model. The WIP helps to control insertion and deletion of words. Note that the true number of words is unknown in the decoding step. Splitting an

[1] For n-gram language models the term $p(w_i|w_{i-1})$ is replaced by $p(w_i|w_{i-n+1}^{i-1})$ where w_{i-n+1}^{i-1} stands for the word sequence $(w_{i-n+1}, \ldots, w_{i-1})$.

[2] In the literature the terms linguistic weight, language weight or, more specifically, language model weight can also be found.

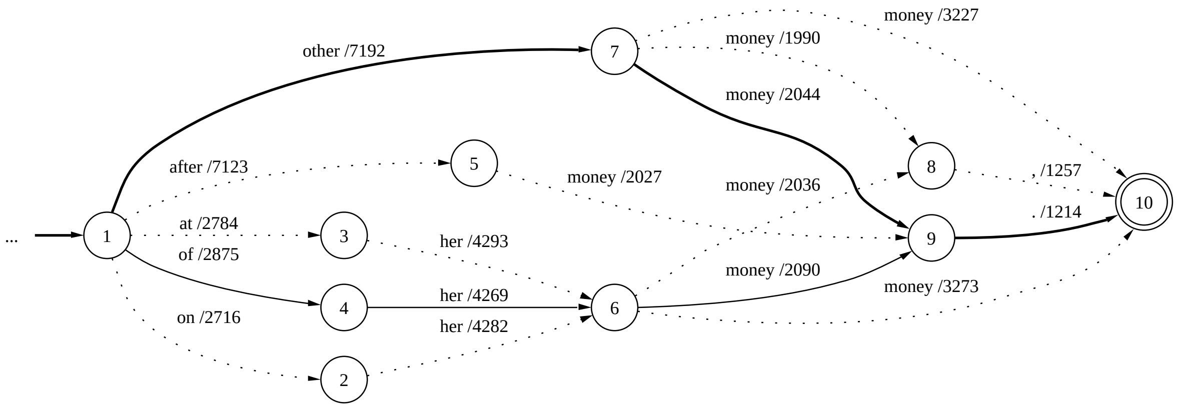

Figure 1. A part of a rescored recognition lattice corresponding to the second line of text in Fig. 2. The bold edges indicate the recognition result and the solid edges represent the correct transcription.

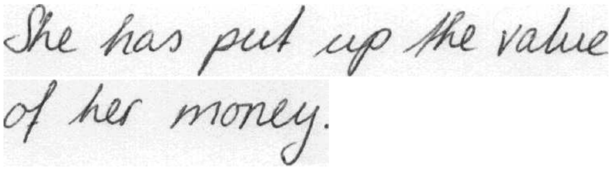
She has put up the value
of her money.

Figure 2. A sample sentence from the IAM Database.

observation sequence X into n words results in a term $n\beta$ to be added to the recognition score for the handwritten sentence. Oversegmentation can be reduced by selecting $\beta \leq 0$ while a choice of $\beta \geq 0$ will decrease the number of undersegmentations[3].

Because there is no exact mathematical model of these two parameters, the optimal values for α and β are normally determined by experiment on a validation set.

2 Related Work

In the domain of handwriting recognition the optimization of the GSF and the WIP has not been addressed before, to the knowledge of the authors. Handwritten text recognition systems which do not include a statistical language model [2, 4, 11] are operating with the parameter set ($\alpha = 0,\ \beta = 0$). Those who do [5, 8, 12], are using ($\alpha = 1,\ \beta = 0$). I.e. the weighting of the output of the language models has not been optimized against the HMM based optical model and no attempt has been made to balance over- and undersegmentation rates.

Some works in the domain of speech recognition investigate the role of the GSF and the WIP [9, 7, 1]. In [9] the optimization of the WIP is addressed. It is suggested to adjust the WIP to control the rate of word insertion and word deletion. A large WIP will reduce the word insertion rate and increase the word deletion rate. A small WIP will have the opposite effect.

According to [1], all current "state-of-the-art" speech recognition systems are not directly applying Bayes' rule; rather, empirical scaling factors are applied to both the language model and the acoustic model[4] output. As an alternative, an approach called mixture of experts is proposed where both the HMM and the statistical language model are treated as mutually exclusive experts. The corresponding scores can then be weighted with priors which estimate the reliability of the experts given a observation sequence X.

In [7] the GSF is shown to be dependent on the sentence length if n-gram language models are used. The factorization of n-gram language models leads to a preference for shorter sentences W. As a consequence, a recognizer often produces sentences which are shorter than the correct answer for large GSF. To compensate for this undersegmentation, negative WIP have been found to produce best results for large GSF.

3 Methodology

Using an HMM based handwritten text recognition system, recognition lattices are produced for handwritten

[3] Please note that positive WIP correspond to negative β values. This comes from the fact that the decoder maximizes the value ϕ_i of Eq. (4).

[4] In the domain of speech recognition the term *acoustic model* is commonly used for the quantity $p(X|W)$.

sentences as shown in Fig. 1 and Fig. 2. These lattices represent part of the search space which has been explored during the Viterbi decoding step and can be seen as directed acyclic graphs. The nodes represent alternative segmentation boundaries and the edges stand for the recognized words. Each edge of a lattice carries the corresponding values $p(X_i|w_i)$ estimated by the HMM as well as the bigram probability $p(w_i|w_{i-1})$ provided by the language model.

A lattice rescoring procedure can then be applied as follows. For a specific parameter pair (α, β) the values of the edges are recombined into single scores using Eq. (4) and the path producing the highest score is printed out as the recognition result. The resulting performance measures, for example the word recognition rate or the word level accuracy[5] , can then be used to jointly optimize the GSF and the WIP on a validation set.

4 Experiments and Results

The HMM based handwritten text recognition system described in [5] is using a linear topology for the character models. This topology was adopted in the current paper. However, the number of states was chosen depending on the individual character [14], and a mixture of eight Gaussians for each state was used, rather than just a single Gaussian as reported in [5]. A writer independent sentence recognition task has been considered using material from the segmented version of the IAM database [6, 13]. The text recognition system was trained on 5,799 images of handwritten text lines containing a total of 39,993 word instances written by 448 different writers. The task lexicon has been closed over the test (validation) set and included 8,819 (8,825) words[6]. For the bigram language model the tagged LOB Corpus [3] was used after excluding all sentences from the test (validation) set. Both validation and test set contain 200 complete sentences each, written by 100 writers which did not contribute to the training set.

For each sentence in the validation set a recognition lattice was computed as described in Sec. 3. The 200 resulting lattices were then rescored using different combinations of the GSF and the WIP. For each combination of the GSF and the WIP the resulting sentence recognition rate, word recognition rate and word level accuracy were measured. In Fig. 3 the resulting sentence recognition rate and word level accuracy are provided where the GSF was sampled in the range of [10,50] and the WIP in the range of [-50,200]. The results for the word recognition rate are not provided since they are very similar to the plot for the word level accuracy.

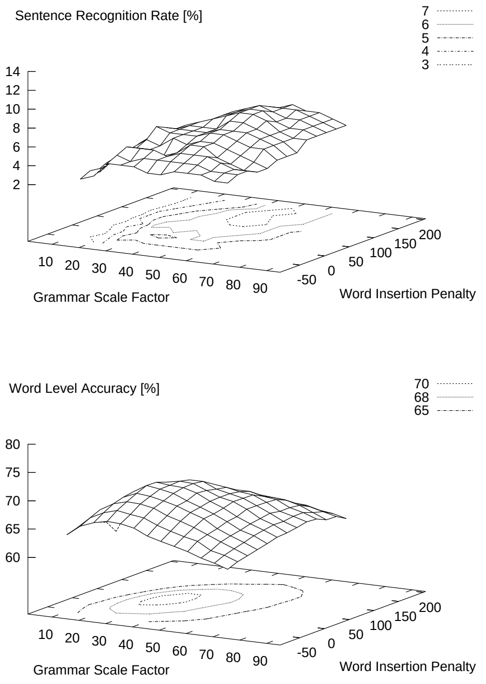

Figure 3. Validation set sentence recognition rate (above) and word level accuracy (below) for different GSF and WIP.

As claimed in [7], it can be seen from Fig. 3 that the GSF and the WIP are not independent of each other. The larger the GSF is, the larger we have to choose the WIP for optimal word level accuracy results. It can further be observed that the best values of the two parameters depend on the performance measure we want to optimize. Parameter which optimize the sentence recognition rate ($\alpha = 45,\ \beta = 75$) differ significantly from those maximizing the word level accuracy ($\alpha = 30,\ \beta = 50$).

Tab. 1 provides three settings of the GSF and the WIP which simulate different recognition systems. For each parameter setting the resulting sentence recognition rate (Sen.), word recognition rate (Wrd.) and word level accuracy (Acc.) are shown for the validation set. For ($\alpha = 0,\ \beta = 0$) a recognition system is assumed which does not use any language model. The second

[5] The word recognition rate is defined as H/N and the word level accuracy as $(H - I)/N$. N represents the number of words in the correct solution, H stands for the number of correctly recognized words and the number of insertions is specified by I.

[6] The closing the lexicon over the test (validation) set ensures that all words of the test (validation) set are contained in the task lexicon.

α	β	Sen.	Wrd.	Acc.
0	0	0.0%	56.7%	46.2%
1	0	0.0%	58.9%	49.4%
30	50	6.0%	73.7%	70.5%

Table 1. Comparison of the validation set performance of the baseline and the optimized systems.

α	β	Sen.	Wrd.	Acc.
0	0	0.5%	60.1%	49.1%
1	0	0.5%	62.4%	52.2%
30	50	10.5%	79.0%	76.3%

Table 2. Comparison of the test set performance of the baseline and the optimized systems.

row ($\alpha = 1,\ \beta = 0$) corresponds to the incorporation of the language model as reported in [5, 8, 12]. The optimized integration given in the third row ($\alpha = 30,\ \beta = 50$) shows the performance gains achieved[7].

Corresponding results for the test set (see Tab. 2) demonstrate that substantial improvements were also achieved on the test set using the GSF and the WIP optimized on the validation set.

5 Conclusions

This paper proposes the use of two parameters, the Grammar Scale Factor (GSF) and the Word Insertion Penalty (WIP) in HMM based handwritten text recognition. The GSF balances the influence of the language model probabilities versus the output of the HMM while the WIP allows to find an optimal trade-off between word insertions and word deletions.

For a writer independent sentence recognition task different combinations of the GSF and the WIP were evaluated to maximize the benefit of a bigram language model in the recognition process. Using the optimal parameter values found on the validation set, the test set word recognition rate could be increased by 16.6% over the previously used language model integration. A corresponding increase of 24.1% was measured for the word level accuracy.

References

[1] H. Bourlard, H. Hermansky, and N. Morgan. Towards increasing speech recognition error rates. *Speech Communication*, 18:205–231, 1996.

[2] R. Bozinovic and S. Srihari. Off-line cursive script word recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(1):68–83, Jan. 1989.

[3] S. Johansson, E. Atwell, R. Garside, and G. Leech. *The Tagged LOB Corpus, Users's Manual.* Norwegian Computing Center for the Humanities, Bergen, Norway, 1986.

[4] E. Kavallieratou, N. Fakotakis, and G. Kokkinakis. An unconstrained handwriting recogntion system. *Int. Journal on Document Analysis and Recognition*, 4:226–242, 2002.

[5] U.-V. Marti and H. Bunke. Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 15:65–90, 2001.

[6] U.-V. Marti and H. Bunke. The IAM-database: an English sentence database for off-line handwriting recognition. *Int. Journal on Document Analysis and Recognition*, 5:39–46, 2002.

[7] A. Ogawa, K. Takeda, and F. Itakura. Balancing acoustic and linguistic probabilities. In *IEEE Conference on Acoustics, Speech and Signal Processing*, pages 181–184, 1998.

[8] F. Perraud, C. Viard-Gaudin, E. Morin, and P.-M. Lallican. N-gram and n-class models for on line handwriting recognition. In *7th Int. Conf. on Document Analysis and Recognition, Edinburgh, Scotland*, volume 2, pages 1053–1057, 2003.

[9] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[10] R. Rosenfeld. Two decades of statistical language modeling: Where do we go from here? *Proc. of the IEEE*, 88:1270–1278, 2000.

[11] A. Senior and A. Robinson. An off-line cursive handwriting recognition system. *IEEE Trans. on Patten Analysis and Machine Intelligence*, 20(3):309–321, Mar. 1998.

[12] A. Vinciarelli, S. Bengio, and H. Bunke. Offline recognition of large vocabulary cursive handwritten text. In *7th Int. Conf. on Document Analysis and Recognition, Edinbourgh, Scotland*, volume 2, pages 1101–1105, 2003.

[13] M. Zimmermann and H. Bunke. Automatic segmentation of the IAM off-line handwritten English text database. In *16th Int. Conf. on Pattern Recognition*, volume 4, pages 35–39, Quebec, Canada, Aug. 2002.

[14] M. Zimmermann and H. Bunke. Hidden Markov model length optimization for handwriting recognition systems. In *8th Int. Workshop on Frontiers in Handwriting Recognition*, pages 369–374, Niagra-on-the-Lake, Canada, Aug. 2002.

[7] Please note that no better sentence recognition rates can be expected given the average length of the sentences (20 words).