

Chapter 5

Active Interaction and Learning in Handwritten Text Transcription

With Contribution Of: Nicolás Serrano, Adrià Giménez, Alberto Sanchís and Alfons Juan.

Contents

5.1	Introduction	119
5.2	Confidence Measures	121
5.3	Adaptation from Partially Supervised Transcriptions	122
5.4	Active Interaction and Active Learning	122
5.5	Balancing Error and Supervision Effort	124
5.6	Experiments	126
5.7	Conclusions	132
	References	132

Computer-assisted systems are being increasingly used in a variety of real-world tasks, though their application to handwritten text transcription in old manuscripts remains largely unexplored. The basic idea explored in this chapter is to follow a sequential, line-by-line transcription of the whole manuscript in which a continuously retrained system interacts with the user to efficiently transcribe each new line. User interaction is expensive in terms of time and cost. Our top priority is to take advantage of these interactions, while trying to reduce them as most as possible.

To this end, we study three different frameworks: (a) improve a recognition system from newly recognized transcriptions via adaptation techniques, using semi-supervised learning techniques; (b) study how to best adapt from limited user supervisions, which is related to active learning; and (c) develop a simple error estimate, which is used to let the user adjust the error in a computer-assisted transcription task. In addition, we test these approaches in the sequential transcription of two old text documents.

5.1 Introduction

Transcription of handwritten text in (old) documents is an important, time-consuming task for digital libraries. It might be carried out by first processing all

document images off-line, and then manually supervising system transcriptions to edit incorrect parts. However, state-of-the-art technologies for automatic page layout analysis, text line detection and handwritten text recognition are still far from perfect [4, 6], and thus post-editing automatically generated output is not clearly better than simply ignoring it.

A more effective approach to transcribe old text documents is to follow an interactive–predictive paradigm in which both, the system is guided by the human supervisor, and the supervisor is assisted by the system to complete the transcription task as efficiently as possible. This computer-assisted transcription approach has been successfully followed in the DEBORA [3] and iDoc [7] research projects, for old-style printed and handwritten text, respectively. In the case of iDoc, a computer-assisted transcription system prototype called GIDOC (Gimp-based Interactive transcription of old text DOCuments) has been developed to provide user-friendly, integrated support for interactive–predictive page layout analysis, text line detection and handwritten text transcription. A detailed description of the GIDOC prototype can be found in Chap. 12.

All works presented in this chapter were performed using GIDOC. As in most of the advanced handwriting recognizers today, it is based on standard speech technology adapted to handwritten text images; that is, HMM-based text image modeling and n -gram language modeling, as introduced in Chap. 2 of this book. The system is trained from manually transcribed text lines during early stages of the transcription task. Then, each new text line image is processed in turn, by first predicting its most likely transcription, and then locating and editing system errors. In order to reduce the effort in locating these errors, GIDOC again resorts to standard speech technology and, in particular, to confidence measures (at word level), which are calculated as posterior word probabilities estimated from word graphs [10]. Recognized words below a given confidence threshold are marked as possible errors, and the decision on how to proceed is left to the user. For instance, if a small number of transcription errors can be tolerated for the sake of efficiency, then the user might validate the system output after only supervising (a few) marked words.

Following previous ideas in the areas of machine translation and speech recognition, a prefix-based interactive–predictive approach is proposed in previous chapters of this book in which the user supervises each new line, in the usual reading order, and corrects the first incorrectly recognized word, if any. The prefix of the current hypothesis is thus validated up to the corrected word, and hence the system updates the current hypothesis by searching for the most probable suffix after the validated prefix. This two-step interactive–predictive process is continued until validation of the whole current hypothesis. It is worth noting that this approach is designed to produce complete, error-free transcriptions of handwritten text. According to the taxonomy outlined in Sect. 1.4.1, this corresponds to a *Passive, Left-to-right* interaction protocol in which the user has to supervise all recognized words. In contrast, the ideas presented in this chapter assume an *Active* interaction protocol which does not need complete supervision (and does not guarantee error-free results).

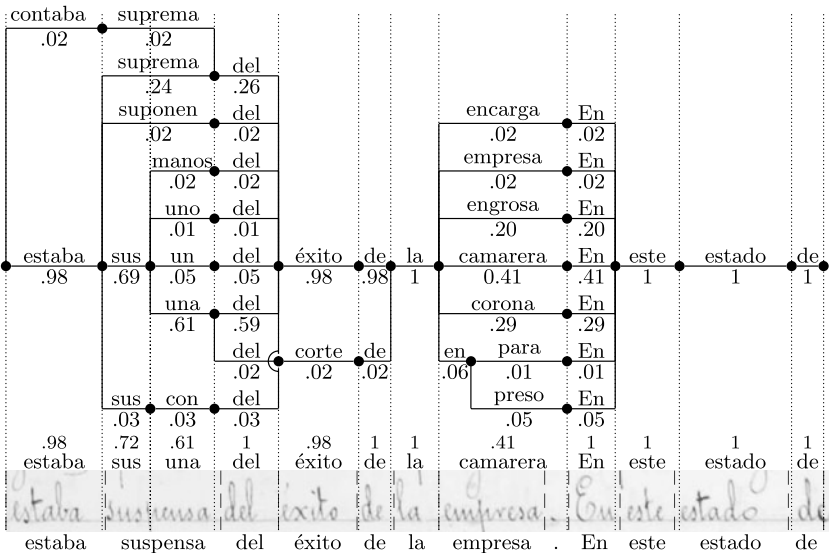


Fig. 5.1 Word-graph example aligned with its corresponding text line image and its recognized and true transcriptions. Each recognized word is labeled (*above*) with its associated confidence measure

The remainder of this chapter is organized as follows: in Sect. 5.2 the use of confidences measures for error locating is explained. Sections 5.3, 5.4 and 5.5 described in detail the main contributions under three different frameworks: adaptive learning, active interaction and learning, and development of simple good error estimate to allow user adjust the error in a computer-assisted transcription task. In the last Sect. 5.6, we test the presented approaches in two real handwriting tasks.

5.2 Confidence Measures

As indicated in the introduction, confidence measures on recognized words are calculated as posterior word probabilities estimated from word graphs. Generally speaking, word graphs are used to represent, in a compact form, large sets of transcription hypotheses with relatively high probability of being correct. See Sect. 1.5.1 for details.

Consider the example in Fig. 5.1, where a small (pruned) word graph is shown aligned with its corresponding text line image and its recognized and true transcriptions.

Each word-graph node is aligned with a discrete point in space, and each edge is labeled with a word (*above*) and its associated posterior probability (*below*). For instance, in Fig. 5.1, the word “sus” has a posterior probability of 0.69 to occur between “estaba” and “un”, and 0.03 to occur between “estaba” and “con”. Note

that all word posteriors sum to 1 at each point in space. Therefore, the posterior probability for a word w to occur at a specific point p is given by the sum of all edges labeled with w that are found at p ; e.g. “sus” has a posterior probability of 0.72 at any point in which the two edges labeled with “sus” are simultaneously found. As discussed in Sect. 1.5.2, the confidence measure of a recognized word is calculated from these point-dependent posteriors, by simply maximizing over all points where it is most likely to occur (Viterbi-aligned). As an example, each recognized word in Fig. 5.1 is labeled (above) with its associated confidence measure. Please see [10] for more details.

5.3 Adaptation from Partially Supervised Transcriptions

In this section, we introduce an interactive transcription framework, where successively produced transcriptions can be used to better adapt image and language models to the task by, for instance, re-training them from the previous and newly acquired transcribed data. However, if transcriptions are only partially supervised, then (hopefully minor) recognition errors may go unnoticed to the user and have a negative effect on model adaptation.

We study this effect as a function of the degree of supervision, i.e. the number of words supervised per line, and as a function of the adaptation strategies used to re-train the system. Concretely, we consider three adaptation strategies: from all data, only from supervised parts, and from high-confidence parts. Re-training from all data is commonly known as unsupervised learning, where a system learns from its own (unmodified) output. Given the user supervisions we can choose to train uniquely from user supervised transcription, as it is typically performed in active learning systems. In the last strategy, re-training from high-confidence parts, we use the best of the two previous approaches. It is inspired in [11], where confidence measures were successfully used to restrict unsupervised learning of acoustic models for large vocabulary continuous speech recognition. It must be noted that, high-confidence parts include both, unsupervised words above certain confidence threshold, and supervised words. Figure 5.2 shows an example of the three strategies.

5.4 Active Interaction and Active Learning

Active learning strategies are being increasingly used in a variety of real-world tasks where user supervision is difficult, time-consuming, or expensive to obtain [9]. Active learning is particularly adequate for *active interaction* protocols, as those studied in this chapter. In interactive transcription of old text documents, the simplest active interaction strategy is to supervise the least confident words of a given recognizer output. Next, active learning consists in adapting the system models by means of these corrected transcriptions, as discussed in the previous section.

0.98 0.72 0.61 1 0.98 1 1 1 1 1 1
 estaba sus una del éxito de la **empresa** . En este estado de
 From all data (Unsupervised)

estaba sus **una** del éxito de la **empresa** . En este estado de
 From user supervised parts

estaba sus una del éxito de la **empresa** . En este estado de
 From high confidence parts ($cm > 0.95$)

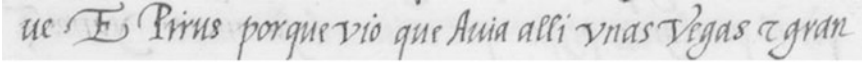
estaba sus una del éxito de la **empresa** . En este estado de

Fig. 5.2 Example showing words (marked in *bold*) which will be used in the next re-training, when using the different adaptation technique. The *first row* shows the recognized line along with its confidence measure (*above*), as well as the words “*empresa*” and “*.*” supervised by the user

In this section, we focus on *active interaction* and explore how it can be used to further enhance system performance. That is, we take advantage of the user feedback, in form of corrected words, to further improve the transcription accuracy. The conventional, non-interactive recognition strategy is improved by letting the system recompute the most probable hypotheses with the constraints imposed by user supervisions. In particular, two strategies, called *iterative* and *delayed*, are studied which differ in the frequency of hypothesis recomputation on the current line.

An application example of the conventional, iterative and delayed strategies is shown in Fig. 5.3, with user supervision limited to three words. The conventional approach leads to the correction of the three words recognized with less confidence (*ras*, *me* and *&*), resulting in a corrected transcription which still contains two incorrectly recognized words (*vn* and *Aguas*). The iterative strategy first asks for the supervision of *ras*, which is substituted by *Pirus*, and then recomputes the most probable hypothesis, where four more recognized words of the previous hypothesis are substituted or deleted (*me*, *Aguas*, *&* and *vn*). The second iteration reduces to substituting *te* for *me*. In the third iteration, the user substitutes *Vegas* for *vengar*, which results in the correct transcription but, somewhat surprisingly, recomputation of the most probable hypothesis ends up with a recognition error (*vna*). The delayed strategy, shown at the bottom of Fig. 5.3, simply amounts to recompute the most probable hypothesis after the conventional (manual) correction of the three words recognized with less confidence. In contrast to the conventional approach, only one recognition error remains in the final transcription (*vengar*).

An important issue regarding the implementation of the iterative and delayed strategies is how to compute a most probable hypothesis compatible with user supervisions and corrections. Following Kristjansson [2], we have implemented a constrained Viterbi decoding algorithm in which the search for the most probable path is constrained to pass through subpaths that conform user supervisions and corrections. More precisely, word scores in supervised segments are set to null for all words but those supervised and possibly corrected.



ue.	E	Pirus	porque	vio	que	Auia	alli	vnas	Vegas	&	gran		
Conventional:													
.5	.7	.9	.4	1	1	1	1	1	.9	.9	1	1	
<u>me</u>	<u>&</u>	E	<u>ras</u>	porque	vio	que	Auia	alli	vnas	<u>vn</u>	<u>Aguas</u>	<u>&</u>	<u>gran</u>
ue.	E	Pirus	porque	vio	que	Auia	alli	vnas	<u>vn</u>	<u>Aguas</u>	<u>&</u>	<u>gran</u>	
Iterative:													
.5	.7	.9	.4	1	1	1	1	1	.9	.9	1	1	
<u>me</u>	<u>&</u>	E	<u>ras</u>	porque	vio	que	Auia	alli	vnas	<u>vn</u>	<u>Aguas</u>	<u>&</u>	<u>gran</u>
.5	1	1	1	.9	1	1	1	.8	.6	1	1		
<u>te</u>	E	Pirus	porque	vio	que	Auia	alli	vnas	<u>vengar</u>	<u>&</u>	<u>gran</u>		
1	1	1	1	.9	1	1	1	.8	.6	1	1		
ue.	E	Pirus	porque	vio	que	Auia	alli	vnas	<u>vengar</u>	<u>&</u>	<u>gran</u>		
ue.	E	Pirus	porque	vio	que	Auia	alli	<u>vn</u>	Vegas	<u>&</u>	<u>gran</u>		
Delayed:													
.5	.7	.9	.4	1	1	1	1	1	.9	.9	1	1	
<u>me</u>	<u>&</u>	E	<u>ras</u>	porque	vio	que	Auia	alli	vnas	<u>vn</u>	<u>Aguas</u>	<u>&</u>	<u>gran</u>
ue.	E	Pirus	porque	vio	que	Auia	alli	vnas	<u>vengar</u>	<u>&</u>	<u>gran</u>		

Fig. 5.3 Application example of the conventional, iterative and delayed strategies for interactive–predictive transcription of a text line image with user supervision limited to three words. Recognized words are labeled above with their associated confidence measures. Supervised words and transcription errors are marked with *plain* and *wavy underlining*, respectively

5.5 Balancing Error and Supervision Effort

In this section, we study how to automatically balance recognition error and supervision effort. Our starting point is a system applying the best adaptation strategy from Sect. 5.3, where we have compared several model adaptation techniques from partially supervised transcriptions. Experiments showed that it is better not to adapt models from all data, but only from high-confidence parts, or just simply from supervised parts. More importantly, it has been shown that a certain degree of supervision is required for model adaptation, although it remains unclear how to adjust it properly. To this end, we propose a simple yet effective method to find an optimal balance between recognition error and supervision effort. The user decides on a maximum tolerance threshold for the recognition error (in non-supervised parts), and the system “*actively*” adjusts the required supervision effort on the basis of an estimate for this error.

Recognition error is measured in terms of Word Error Rate (WER); that is, as the average number of elementary editing operations needed to produce a reference (correctly transcribed) word from recognized words. Given a collection of reference-recognized transcription pairs, its WER may be simply expressed as

$$\text{WER} = \frac{E}{N},$$

where E is the total number of editing operations required to transform recognized transcriptions into their corresponding references, and N is the total number of reference words. In this work, however, we need to decompose these three variables additively, as

$$\begin{aligned} \text{WER} &= \widehat{\text{WER}}^+ + \text{WER}^-, \\ E &= E^+ + E^- \quad \text{and} \quad N = N^+ + N^-, \end{aligned}$$

where the superscripts $^+$ and $^-$ denote supervised and unsupervised parts, respectively, and thus

$$\text{WER}^+ = \frac{E^+}{N} \quad \text{and} \quad \text{WER}^- = \frac{E^-}{N}.$$

In order to balance error and supervision effort, we propose the system to ask for supervision effort only when WER^- becomes greater than a given, maximum tolerance threshold, say WER^* . However, as we do not know the values of E^- and N^- , they have to be estimated from the available data. A reasonable estimate for N^- is simply

$$\hat{N}^- = \frac{N^+}{R^+} R^-,$$

where R^+ and R^- denote the number of recognized words in the supervised and unsupervised parts, respectively. Similarly, a reasonable estimate for E^- is

$$\hat{E}^- = \frac{E^+}{R^+} R^-$$

and thus the desired estimate for WER^- is

$$\widehat{\text{WER}}^- = \frac{\frac{E^+}{R^+} R^-}{N^+ + \frac{N^+}{R^+} R^-}.$$

Each recognized word will be accepted without supervision if it does not lead to a $\widehat{\text{WER}}^-$ estimate greater than WER^* .

Note that the above estimate for WER^- is pessimistic, since it assumes that, on average, correction of unsupervised parts requires similar editing effort to that required for supervised parts. However, the user is asked to supervise recognized words in increasing order of confidence, and hence unsupervised parts should require less correction effort. In order to better estimate WER^- , we may group recognized words by their level of confidence c , from 1 to a certain maximum level C , and compute a c -dependent estimate for E as above,

$$\hat{E}_c^- = \frac{E_c^+}{R_c^+} R_c^-$$

where E_c^+ , R_c^+ and R_c^- are c -dependent versions of E^+ , R^+ and R^- , respectively. The global estimate for E is obtained by simply summing these c -dependent estimates,

$$\hat{E}^- = \sum_{c=1}^C \hat{E}_c^-$$

and, therefore, the estimate for WER^- becomes

$$\widehat{\text{WER}}^- = \frac{\sum_{c=1}^C \frac{E_c^+}{R_c^+} R_c^-}{N^+ + \frac{N^+}{R^+} R^-}$$

which reduces to the previous, pessimistic estimate when only a single confidence level is considered ($C = 1$).

5.6 Experiments

In the following sections, the active learning and interactive transcription strategies described are applied in two real handwritten tasks: GERMANA and RODRIGO.

5.6.1 User Interaction Model

In order to validate our interactive transcription techniques, we need to perform a high number of experiments. As our experiments require from user supervision, dealing with real users would be impossible because of time and cost. In this section, we propose a simple yet realistic user interaction model to simulate user actions at different degrees of supervision. The degree of supervision is modeled as the (maximum) number of recognized words (per line) that are supervised: 0 (unsupervised), 1, ..., ∞ (fully supervised). It is assumed that recognized words are supervised in non-decreasing order of confidence.

In order to predict the user actions associated with each word supervision, we first compute a minimum edit (Levenshtein) distance path between the recognized and true transcriptions of a given text line. For instance, the example text line image in Fig. 5.1 is also used in Fig. 5.4 to show an example of minimum edit distance path between its recognized and true transcriptions. As usual, three elementary editing operations are considered: substitution (of a recognized word by a different word), deletion (of a recognized word) and insertion (of a missing word in the recognized transcription). Substitutions and deletions are directly assigned to their corresponding recognized words. In Fig. 5.4, for instance, there is a substitution assigned to “sus”, a deletion assigned to “una”, and a second substitution that corresponds to “camarera”. Insertions, however, have not direct assignments to recognized words and, hence, it is not straightforward to predict when they are carried out by the user.

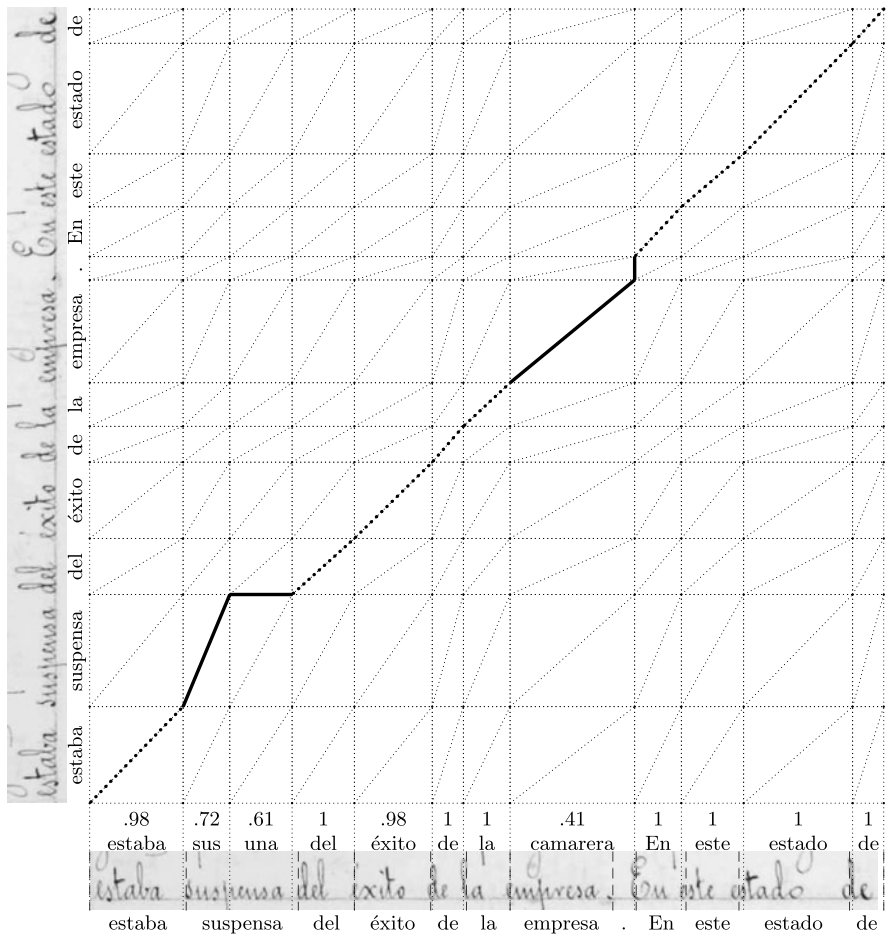


Fig. 5.4 Example of minimum edit distance path between the recognized and true transcriptions of a text line image

To this end, we first compute the Viterbi segmentations of the text line image from the true and recognized transcriptions. Given a word to be inserted, it is assigned to the recognized word whose Viterbi segment covers most part of its true Viterbi segment. For instance, in Fig. 5.4, the period is completely covered by “camarera”, and thus its insertion is assumed to be done when “camarera” is supervised.

5.6.2 Sequential Transcription Tasks

Experiments were carried out on two datasets recently introduced: GERMANA [5] and RODRIGO [8]. GERMANA is the result of digitizing and annotating a 764-

Table 5.1 Statistics of GERMANA and RODRIGO. Singletons corresponds to words appearing once in the document. Perplexity drawn from a bigram language model in a ten-fold validation

	GERMANA	RODRIGO
Pages	764	853
Lines	20529	20357
Running words (K)	217	232
Lexicon size (K)	27.1	17.3
Singletons (%)	57.4	54.4
Character set size	115	115
Perplexity	290	166

page Spanish manuscript from 1891, in which most pages only contain nearly calligraphed text written on ruled sheets of well-separated lines. The example shown in Fig. 5.1 contains a text line image from GERMANA. GERMANA is solely written in Spanish up to p. 180, but then it includes many parts written in languages other than Spanish. RODRIGO is similar to GERMANA both, in size and page layout. However, it comes from a much older manuscript, from 1545, and it is completely written in Spanish. As can be seen in text line image shown at the top of Fig. 5.3, which was extracted from p. 65, the writing style has clear Gothic influences. Some basic statistics of GERMANA and RODRIGO are provided in Table 5.1.

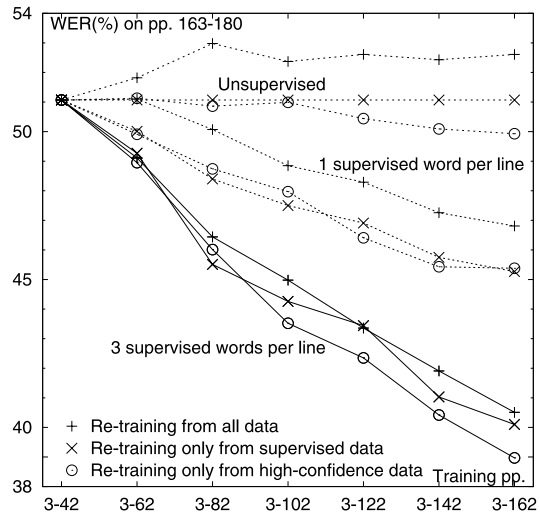
5.6.3 Adaptation from Partially Supervised Transcriptions

Due to its sequential book structure, the very basic task on GERMANA is to transcribe it from the beginning to the end, though here we only consider its transcription up to p. 180. Starting from p. 3, we divided GERMANA into nine consecutive blocks of 20 pages each (18 in block 9). The first two blocks (pp. 3–42) were used to train initial image and language models from fully supervised transcriptions. Then, from block 3 to 8, each new block was recognized, partially supervised and added to the training set built from its preceding blocks.

As it has been said in Sect. 5.3, we perform the sequential transcription of GERMANA as function of: the degree of supervision and the adaptation technique used. We considered three degrees of supervision: zero (unsupervised), one and three supervised words per line; and the three adaptation (re-training) strategies: from all data, only from high-confidence parts, and only from supervised parts. The results are shown in Fig. 5.5 in terms of Word Error Rate (WER) on block 9 (pp. 163–180).

From the results in Fig. 5.5, it becomes clear that baseline models can be improved by adaptation from partially supervised transcriptions, though a certain degree of supervision is required to obtain significant improvements. In particular, supervision of three words per line leads to a reduction of more than a 10% of WER with respect to unsupervised learning (baseline models), though there is still room for improvement since full supervision achieves a further reduction of 5% (34%). The adaptation strategy, on the other hand, has a relatively minor effect on

Fig. 5.5 Test-set Word Error Rate (WER) on GERMANA as a function of the training set size (in pages), for varying degrees of supervision (supervised words per line)



the results. Nevertheless, it seems better not to re-train from all data, but only from high-confidence parts, or just simply from supervised parts.

Apart from the above experiment on GERMANA, we did a similar experiment on the well-known IAM dataset, using a standard partition into a training, validation and test sets [1]. The training set was further divided into three subsets; the first one was used to train initial models, while the other two were recognized, partially supervised (four words per line) and added to the training set. The results obtained in terms of test-set WER are: 42.6%, using only the first subset; 42.8%, after adding the second subset; and 42.0%, using also the third subset. In contrast to GERMANA, there is no significant reduction in terms of WER after adding partially supervised data to the training set. We think that this result is due to the more complex nature of the IAM task.

5.6.4 Active Interaction and Learning

In this section, we describe the experiments done to test the active learning strategies referred in Sect. 5.4. In this set of experiments, we used the best system from the previous experiment. Again, the quality of the successively produced models was measured in terms of WER on block 9, and it is shown in Fig. 5.6 (left). Full supervision (∞) and the conventional strategy (C) are compared with the two strategies discussed; that is, iterative (I) and delayed (D). The C, I and D strategies were limited to three supervised words per line, which is not too much since, on average, text lines are of 11 words approximately.

Experiments similar to those previously described were also carried out on RODRIGO. The 20K lines of RODRIGO were divided into 20 consecutive blocks of 1 000 lines approximately, except for the first 1 000 lines, which were divided into

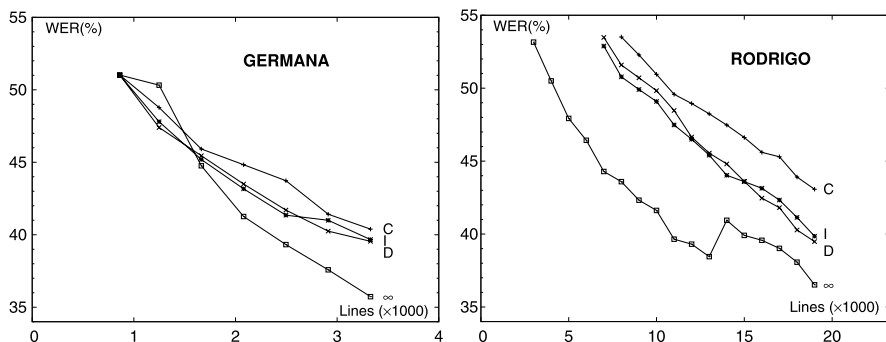


Fig. 5.6 Word Error Rate (WER) on the last block of lines, as a function of the number of training lines, for full supervision (∞), and partial supervision (of three words per line), using three active learning strategies: conventional (C), iterative (I), and delayed (D). *Left: GERMANA. Right: RODRIGO*

the line blocks 1–100, 101–200, 201–500 and 501–1 000. The results are also shown in Fig. 5.6 (right).

From the results in Fig. 5.6, it becomes clear that the proposed iterative and delayed strategies are better than the basic, conventional approach. In the case of RODRIGO, conventional supervision of three words per line results in a WER of 43.1%, which is 6.6 points above full supervision (36.5%). By contrast, the iterative and delayed strategies are only 3.3 and 3.0 points above, respectively. That is, the increase of WER due to supervising only three words per line is halved by using the proposed strategies. Moreover, it is worth noting that this increase of 3 points over a WER of 36.5% is just a small degradation in terms of WER, as compared with the considerable user effort reduction achieved by only supervising three out of 11 words per line. On the other hand, it seems that the iterative and delayed strategies produce nearly identical results, though this should be further explored by also considering the effect of varying the supervision degree (number of supervised words per line).

In the case of GERMANA, the iterative and delayed strategies also provide better results than the conventional approach, though the WER improvements are more moderate. This might be due to the fact that GERMANA models are produced from training sets much smaller than those used for RODRIGO. Note that GERMANA is easier to recognize than RODRIGO, since WER results similar to those obtained on RODRIGO are achieved from much less training lines.

5.6.5 Balancing User Effort and Recognition Error

Perfect transcription of old text documents is not always mandatory. Transcriptions containing a few number of errors are perfectly readable and can be easily obtained using a computer-assisted system. In Sect. 5.5 we introduce a simple yet effective

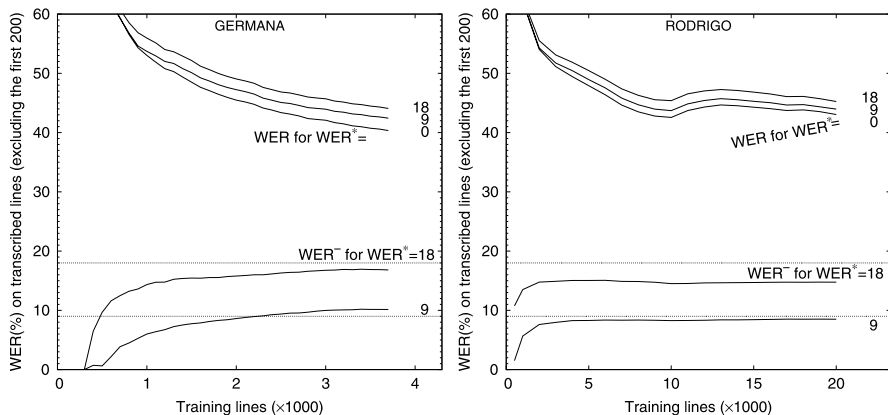


Fig. 5.7 Word Error Rate (WER) on transcribed lines (excluding the first 200), as a function of the (number of) training lines, for varying tolerance thresholds on the recognition error (in unsupervised parts). *Left*: GERMANA dataset. *Right*: RODRIGO dataset

method to balance error and user effort. Here, we consider the transcription under three tolerance thresholds on the recognition error (in unsupervised parts): 0% (fully supervised), 9% (one recognition error per line, on average) and 18%.

In this case, we divided GERMANA into consecutive blocks of 100 lines each (37 blocks). The first two blocks were used to train initial image and language models from fully supervised transcriptions. Then, from block 3 to 37, each new block was recognized, partially supervised as discussed in Sect. 5.5 for $C = 4$ confidence levels, and added to the previous training set. The first three confidence levels correspond, respectively, to the first three words in each line that were recognized with smaller confidence; the remaining recognized words were all grouped into the fourth level. Re-training of image and language models was carried out from only high-confidence parts [7]. The results are shown in Fig. 5.7 (left) in terms of WER on transcribed lines (excluding the first 200).

From the results in Fig. 5.7 (left), it becomes clear that the proposed balancing method takes full advantage of the allowed tolerance to reduce the supervision effort. Moreover, the total WER of the system trained with partial transcriptions does not deviate significantly from that of the fully supervised system. The average user effort reduction ranges from 29% (for $WER^* = 9\%$) to 49% (for $WER^* = 18\%$). That is, if one recognition error per line is allowed for on average ($WER^* = 9\%$), then the user will save 29% of the supervision actions that are required in the case of a fully supervised system. Here, supervision actions refers to elementary editing operations, and also to check that a correctly recognized word is certainly correct.

In order to better assess the proposed method, a larger experiment was also conducted on RODRIGO, which was divided into blocks of 1 000 lines each, except for the first 1 000 lines, which were divided into the line blocks 1–100, 101–200, 201–500 and 501–1 000. The experiment and results, shown in Fig. 5.7 (right), are analogous to those described above for GERMANA.

Although the results presented in Fig. 5.7 are quite satisfactory, we have observed that the proposed balancing method does not clearly favor supervision of low confidence words over those recognized with high confidence. We think that this is mainly due to the fact that it works on a word-by-word basis and, in order to decide whether a given word has to be supervised or not, its contribution to the current estimate of WER^- is not as important as the closeness of this estimate to WER^* . We think that this behavior can be alleviated by using more confidence levels or, more directly, by working on a line-by-line basis. That is, by first assuming that all balancing error recognized words in a line are not supervised, and then supervising words in increasing order of confidence while the current estimate of WER^- is above WER^* .

5.7 Conclusions

In this chapter we described three different frameworks to deal with the interactive transcription process of handwritten documents, where the recognizer output is partially supervised. The basic idea is to assist the user in the transcription process, while keeping his interactions as low as possible. It has been shown that, a system can be trained from partially (and possibly erroneous) supervised transcription, while achieving similar results to a fully supervised trained system. We showed that user interaction can be used to further improve the current transcription, constraining the current hypothesis search space. Lastly, we created a framework that allows the user to adjust the error in exchange of user effort. Experiments were performed on two real transcription tasks, GERMANA and RODRIGO, showing the effectiveness of the proposed frameworks.

References

1. Bertolami, R., & Bunke, H. (2008). Hidden Markov model-based ensemble methods for offline handwritten text line recognition. *Pattern Recognition* 41, 3452–3460.
2. Kristjansson, T., Culotta, A., Viola, P., & McCallum, A. (2004). Interactive information extraction with constrained conditional random fields. In *Proceedings of the 19th national conference on artificial intelligence (AAAI 2004)* (pp. 412–418), San Jose, CA, USA.
3. Le Bourgeois, F., & Emptoz, H. (2007). DEBORA: Digital AccEss to BOoks of the RenAissance. *International Journal on Document Analysis and Recognition*, 9, 193–221.
4. Likforman-Sulem, L., Zahour, A., & Taconet, B. (2007). Text line segmentation of historical documents: a survey. *International Journal on Document Analysis and Recognition*, 9, 123–138.
5. Pérez, D., Tarazón, L., Serrano, N., Castro, F., Ramos-Terrades, O., & Juan, A. (2009). The GERMANA database. In *Proceedings of the 10th international conference on document analysis and recognition (ICDAR 2009)* (pp. 301–305), Barcelona, Spain.
6. Plötz, T., & Fink, G. A. (2009). Markov models for offline handwriting recognition: a survey. *International Journal on Document Analysis and Recognition*, 12, 269–298.

7. Serrano, N., Pérez, D., Sanchis, A., & Juan, A. (2009). Adaptation from partially supervised handwritten text transcriptions. In *Proceedings of the 11th international conference on multimodal interfaces and the 6th workshop on machine learning for multimodal interaction (ICMI-MLMI 2009)* (pp. 289–292), Cambridge, MA, USA.
8. Serrano, N., Castro, F., & Juan, A. (2010). The RODRIGO database. In *Proceedings of the 7th international conference on language resources and evaluation (LREC 2010)* (pp. 2709–2712), Valletta, Malta.
9. Settles, B. (2009). *Active learning literature survey* (Computer Sciences Technical Report No. 1648). University of Wisconsin-Madison.
10. Tarazón, L., Pérez, D., Serrano, N., Alabau, V., Ramos-Terrades, O., Sanchis, A., & Juan, A. (2009). Confidence measures for error correction in interactive transcription of handwritten text. In *Proceedings of the 15th international conference on image analysis and processing (ICIAP 2009)* (pp. 567–574), Vietri sul Mare, Italy.
11. Wessel, F., & Ney, H. (2005). Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(1), 23–31.