Del Agua Teba, MA.; Serrano Martinez Santos, N.; Juan Císcar, A. (2011). Language identification for interactive handwriting transcription of multilingual documents. En Pattern Recognition and Image Analysis. Springer Verlag (Germany). 6669:596-603. doi:10.1007/978-3-642-21257-4_74.

Language Identification for Interactive Handwriting Transcription of Multilingual Documents

Miguel A. del Agua, Nicolás Serrano and Alfons Juan

DSIC/ITI, Universitat Politècnica de València
{mdelagua,nserrano,ajuan}@dsic.upv.es

Abstract. An effective approach to handwriting transcription of (old) documents is to follow a sequential, line-by-line transcription of the whole document, in which a continuously retrained system interacts with the user. In the case of multilingual documents, however, a minor yet important issue for this interactive approach is to first identify the language of the current text line image to be transcribed. In this paper, we propose a probabilistic framework and three techniques for this purpose. Empirical results are reported on an entire 764-page multilingual document for which previous empirical tests were limited to its first 180 pages, written only in Spanish.

Keywords: Language Identification, Interactive Handwriting Transcription, Multilingual Documents

1 Introduction

Transcription of handwritten text in (old) documents is an important, time-consuming task for digital libraries. However, automated techniques for document image analysis and handwriting recognition are still far from perfect [4], and thus post-editing automatically generated output is not clearly better than simply ignoring it. Instead, a more effective approach is to follow a sequential, line-by-line transcription of the whole document, in which a continuously retrained system interacts with the user. In this way, the main task of the user is to (partially) supervise and correct, if needed, each new line transcription hypothesis of the system. This interactive handwriting transcription approach, also extended to interactive layout analysis and line detection, has been implemented in an open source tool called *Gimp-based Interactive transcription of old text DOCuments* (GIDOC) [8]. Using this tool, we have recently studied how to better adapt models from partially supervised transcriptions [6], how to properly balance error and supervision effort [7], as well as different active learning strategies to interact with the user on each new system hypothesis [5].

In the case of multilingual documents, however, a minor yet important issue for interactive transcription of a text line image (or an image block) is to first identify its corresponding language. A good example of multilingual document

is the GERMANA database [3]. GERMANA is the result of digitizing and annotating a 764-page, single-author Spanish manuscript from 1891, solely written in Spanish up to page 180, but then also written in five other languages, especially Catalan and Latin. For simplicity, to avoid dealing with multiple languages, we limited ourselves to the first 180 pages of GERMANA in the empirical tests of the studies cited above.

To our knowledge, however, language identification for interactive transcription of multilingual documents remains unexplored. Indeed, conventional (non-interactive) script and language identification in handwritten documents is still in its early stage of research [2]. In this paper, after a brief review of GIDOC, we propose a probabilistic framework and three techniques for language identification in interactive transcription of multilingual documents (Section 3). In Section 4, empirical results are reported on the whole GERMANA manuscript. Conclusions drawn and future work are summarized in Section 5.

2 GIDOC overview

GIDOC is a first attempt to provide user-friendly, integrated support for interactive-predictive page layout analysis, text line detection and handwritten text transcription [8]. It is built as a set of plug-ins for the well-known GNU Image Manipulation Program (GIMP), which has many image processing features already incorporated and, what is more important, a high-end user interface for image manipulation. To run GIDOC, we must first run GIMP and open a document image. GIMP will come up with its high-end user interface, which is often configured to only show the main toolbox and an image window. GIDOC can be accessed from the menubar of the image window (Fig. 1).

As shown in Fig. 1, the GIDOC menu includes six entries, though here only the last one, *Transcription,* is briefly described (see [8] for a more detailed description). The *Transcription* entry opens an interactive transcription dialog (also shown in Fig. 1), which consists of two main sections: the image section, in the middle part, and the transcription section, in the bottom part. A number of text line images are displayed in the image section together with their transcriptions, if available, in separate editable text boxes within the transcription section. The *current* line to be transcribed is selected by placing the edit cursor in the appropriate editable box. Its corresponding baseline is emphasized (in blue color) and, whenever possible, GIDOC shifts line images and their transcriptions so as to display the current line in the central part of both the image and transcription sections. It is assumed that the user transcribes or supervises text lines, from top to bottom, by entering text and moving the edit cursor with the arrow keys or the mouse.

Note that each editable text box has a button attached to its left, which is labeled with its corresponding line number. By clicking on it, its associated line image is extracted, preprocessed, transformed into a sequence of feature vectors, and Viterbi-decoded using character HMMs and a language model previously trained (see [8] for details on preprocessing, feature extraction, HMM-based im-

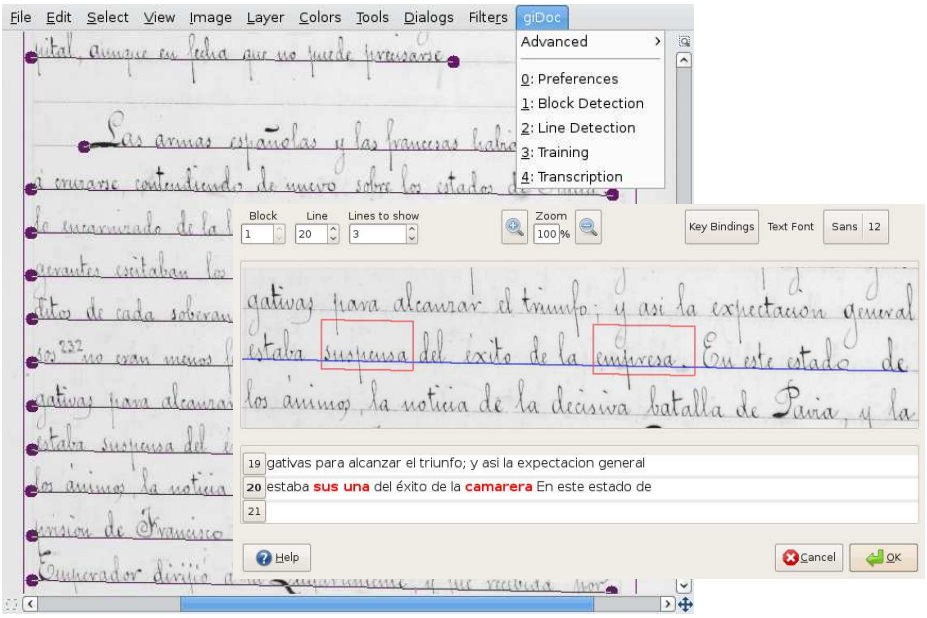

Fig. 1. Interactive Transcription dialog over an image window showing GIDOC menu.

age modeling and language modeling in GIDOC). As shown in Fig. 1, words in the current line for which the system is not highly confident are emphasized (in red) in both the image and transcription sections.

3 Probabilistic framework

Let l be the number of the current text line image to be transcribed, and let x_l be its corresponding sequence of feature vectors. The task of our system is to predict first its (unknown) language identification label, c_l, and then its transcription, w_l. We assume that all preceding lines have been already annotated in terms of language labels, c_1^{l-1}, and transcriptions, w_1^{l-1}. By application of the Bayes decision rule, the minimum-error system prediction for c_l is:

$$c_l^*(x_l, c_1^{l-1}) = \underset{\tilde{c}_l}{\operatorname{argmax}}\ p(\tilde{c}_l \mid x_l, c_1^{l-1}) \tag{1}$$

$$= \underset{\tilde{c}_l}{\operatorname{argmax}}\ p(\tilde{c}_l \mid c_1^{l-1})\, p(x_l \mid \tilde{c}_l) \tag{2}$$

$$= \underset{\tilde{c}_l}{\operatorname{argmax}}\ p(\tilde{c}_l \mid c_1^{l-1}) \sum_{\tilde{w}_l \in W(\tilde{c}_l)} p(\tilde{w}_l \mid \tilde{c}_l)\, p(x_l \mid \tilde{c}_l, \tilde{w}_l) \tag{3}$$

$$\approx \underset{\tilde{c}_l}{\operatorname{argmax}}\ p(\tilde{c}_l \mid c_1^{l-1}) \max_{\tilde{w}_l \in W(\tilde{c}_l)} p(\tilde{w}_l \mid \tilde{c}_l)\, p(x_l \mid \tilde{c}_l, \tilde{w}_l) \tag{4}$$

where: in Eq. (2), it is assumed that x_l is conditionally independent of all preceding line language labels, c_1^{l-1}, given the current line language label, $\tilde{c}_l$; in Eq. (3), we consider all possible transcriptions for the language $\tilde{c}_l$, $W(\tilde{c}_l)$; and, in Eq. (4), the Viterbi (maximum) approximation to the sum in Eq. (3) is applied to only consider the most likely transcription.

The decision rule (4) requires a *language identification model* for $p(\tilde{c}_l \mid c_1^{l-1})$ and, for each possible language $\tilde{c}_l$, a $\tilde{c}_l$-dependent *language (transcription) model* for $p(\tilde{w}_l \mid \tilde{c}_l)$ and a $\tilde{c}_l$-dependent *image model* for $p(x_l \mid \tilde{c}_l, \tilde{w}_l)$. As done in language modeling for monolingual documents, the language models in the multilingual case, both for identification and transcription, can be implemented in terms of *n-gram language models* [8]. Those for language-dependent transcription can be implemented as usual in the monolingual case though, in our case, each language $\tilde{c}_l$ will have its own n-gram language model, trained only from available transcriptions labeled with $\tilde{c}_l$. Regarding the n-gram language identification model, $p(\tilde{c}_l \mid c_1^{l-1})$, in this paper we propose and compare three rather simple techniques:

1. A *bigram* model estimated by relative frequency counts:

$$\hat{p}(\tilde{c}_l \mid c_{l-1}) = \frac{N(c_{l-1}\tilde{c}_l)}{N(c_{l-1})} \tag{5}$$

2. A *unigram* model also estimated by relative frequency counts:

$$\hat{p}(\tilde{c}_l \mid c_{l-1}) = \frac{N(\tilde{c}_l)}{l-1} \tag{6}$$

3. And a *"copy the preceding label" (CPL)* bigram model:

$$\hat{p}(\tilde{c}_l \mid c_{l-1}) = \begin{cases} 1 & \tilde{c}_l = c_{l-1} \\ 0 & \tilde{c}_l \neq c_{l-1} \end{cases} \tag{7}$$

where $N(\cdot)$ denotes the number of occurrences of a given event in the preceding lines, such as the bigram $c_{l-1}\tilde{c}_l$ or the unigram $\tilde{c}_l$. Note that (5) and, especially (7), assume that consecutive text lines are usually written in the same language. This is not necessarily true though, in the kind of manuscripts (applications) we have in mind (e.g GERMANA), it is a reasonable assumption.

Also as in the monolingual case, the *image models* for the different languages can be implemented in terms of *character HMMs* [8]. Moreover, if only a single script is used for all the languages considered (e.g. Latin), then a single, shared image model for all of them might produce better recognition results than a separate, independent model for each language. Clearly, this can be particularly true for infrequent languages.

Finally, it is often useful in practice to introduce scaling parameters in the decision rule so as to empirically adjust the contribution of the different models involved. In our case, the decision rule given in Eq. (4) can be rewritten as

$$c_l^*(x_l, c_1^{l-1}) \approx \operatorname*{argmax}_{\tilde{c}_l} \; p(\tilde{c}_l \mid c_1^{l-1})^{\beta} \max_{\tilde{w}_l \in W(\tilde{c}_l)} p(\tilde{w}_l \mid \tilde{c}_l)^{\alpha_{\tilde{c}_l}} \, p(x_l \mid \tilde{c}_l, \tilde{w}_l) \tag{8}$$

where we have introduced an *Identification Scale Factor (ISF)* β and, for each language $\tilde{c}_l$, a language-dependent *Grammar Scale Factor (GSF)* $\alpha_{\tilde{c}_l}$. Obviously, Eq. (8) does not differ from Eq. (4) when all these scaling parameters are simply set to unity. In the experiments reported below, these parameters will be adapted from a validation set.

4 Experiments

As indicated in the introduction, the experiments reported here were carried out on a multilingual, single-author manuscript from 1891 known as GERMANA database [3]. Our main goal is to empirically compare the three language identification techniques described in the preceding section. Moreover, we provide recognition results on the complete manuscript, which is also worth noting since previous results on GERMANA have been limited to its first 180 pages, solely written in Spanish. The complete manuscript, which comprises a total of 764 pages, includes five other languages, most notably Catalan and Latin.

Some basic yet precise statistics of GERMANA are given in Table 1. In terms of running words, Spanish comprises about 81% of the document, followed by Catalan (12%) and Latin (4%), while the other three languages only account for less than a 3%. Similar percentages also apply for the number of lines. In terms of lexicons, it is worth noting that Spanish and, to a lesser extent, Catalan and Latin, have lexicons comparable in size to standard databases [3]. Also note that the sum of individual lexicon sizes (29.9K) is smaller than the size of the global lexicon (27.1K). This is due to presence of words common to different languages, such as common words in Spanish and Catalan. On the other hand, singletons, that is, words occurring only once, account for most words in each lexicon ($55\% - 71\%$). It goes without saying that, as usual, language modelling is a difficult task. To be more precise, in Table 1 we have included the global perplexity and the perplexity of each language, as given by a bigram model on a 10-fold cross-validation experiment.

Language	Lines	Words	Lexicon	Singletons	Perplexity
All	20000	217K	27.1K	57.4%	289.8±17.0
Spanish	80.9%	81.4%	19.9K	55.6%	238.1±27.7
Catalan	11.8%	12.4%	4.6K	63.2%	112.9±61.6
Latin	4.6%	3.8%	3.4K	69.2%	211.1±51.3
French	1.3%	1.4%	1.1K	71.1%	88.3±21.0
German	1.1%	0.7%	0.6K	52.7%	92.1±29.2
Italian	0.3%	0.3%	0.3K	67.3%	63.3±14.4

Table 1. Basic statistics of GERMANA.

We divided GERMANA into 40 blocks of 500 lines each. The first block was fully transcribed and an initial system was trained from it. Then, from block 2 to 40, each new block was recognized by the system trained from all

preceding blocks, with the last block being also used as a validation set for parameter adaptation. It is worth noting that, after recognition of each block, the user supervises and, if needed, corrects both, language identification labels and transcriptions.

As a baseline, we first tried a conventional, *monolingual* system, that is, a system assuming that all lines come from a single language, and thus only requiring one language (transcription) model and one image model. On the other hand, we tried four *multilingual* systems which only differ in the way they identify the language of the current line: *supervised* (manually given), *bigram* (using Eq. (5)), *unigram* (using Eq. (6)) and *CPL* (using Eq. (7)). Clearly, in all these four systems, a different language (transcription) model was required for each of the 6 languages in GERMANA. However, as suggested at the end of the preceding section, a single, shared image model was used instead of a separate, independent image model for each language in GERMANA. The results are plotted in Fig. 2, in terms of *Word Error Rate* (WER) of the recognized text up to the current line.

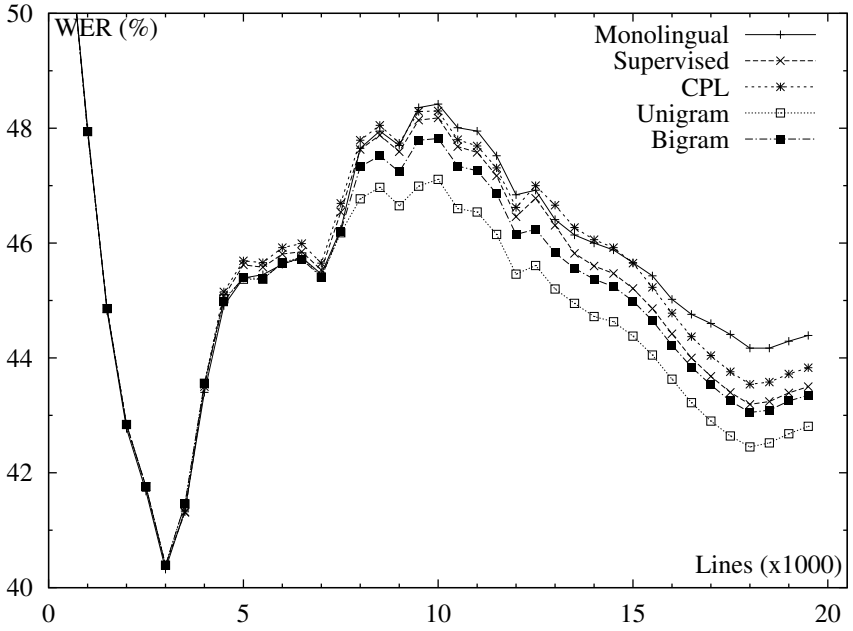

Fig. 2. WER in GERMANA as a function of the number of recognized lines.

As expected, the multilingual systems achieve better results than the monolingual system. Also as expected, the correct language identification label (supervised) produces better results than an automatic, error-prone technique such as CPL. Surprisingly, however, the unigram and, to a lesser extent, the bigram

identification techniques achieve better results than manual supervision. In other words, it is sometimes preferable not to use the correct, but probably poorly-trained language (transcription) model, and use instead a well-trained model for a different yet close language (e.g. Catalan and Spanish). On the other hand, it can be also observed that there are certain blocks at which the WER curve abruptly changes from a (smooth) decreasing tendency to a rapid increase. This was studied carefully in [1] by decomposing the (total) WER curve into its corresponding language-dependent WER curves. It was found that these abrupt changes are due to the occurrence of text from previously unseen languages, most notably Catalan (from line 3500) and Latin (from line 4000).

Although optimal (supervised) language identification does not necessarily lead to better recognition results than those obtained with suboptimal (imperfect) identification techniques, it is still important to have an identification technique of minimal error, maybe to just minimize user effort while correcting identification errors. Table 2 shows the Identification Error Rate (IER) of the proposed techniques for all and each language in GERMANA and both, in absolute and relative terms.

		IER (absolute)			IER (%)		
Language	Lines	2-gram	1-gram	CPL	2-gram	1-gram	CPL
All	19500	1290	2183	488	6.6	11.2	2.5
Spanish	15725	243	312	224	1.5	2.0	1.4
Catalan	2414	534	1136	181	22.1	47.1	7.5
Latin	951	255	409	49	26.8	43.0	5.2
French	266	116	182	31	43.6	68.4	11.7
German	76	74	76	2	97.4	100.0	2.6
Italian	68	68	68	1	100.0	100.0	1.5

Table 2. Identification Error Rate (IER) on GERMANA for the techniques proposed.

From the results in Table 2, it becomes clear that the simplest technique, CPL, is also the most accurate. It achieves an IER of 2.5%, that is, on average, only 3 identified labels out of 100 need to be corrected by the user. In contrast, the 1-gram and 2-gram techniques clearly fail in identifying languages other than Spanish. This might be due to the fact that scaling parameters were adapted to minimize the WER instead of the IER and, indeed, these techniques provided better results than CPL in terms of WER.

5 Conclusions

We have proposed a probabilistic framework and three techniques for language identification in interactive transcription of multilingual documents. These techniques are called the bigram, unigram and CPL-bigram models. They have been empirically compared on the whole GERMANA database, a 764-page, single-author manuscript from 1891 written in six different Latin languages, mainly

Spanish. According to the empirical results, the simplest technique, that is, the "copy the preceding label" (CPL) bigram model is also the most accurate.

Acknowledgements Work supported by the EC (FEDER, FSE), the Spanish Government (MICINN, MITyC, "Plan E", under grants MIPRCV "Consolider Ingenio 2010" , MITTRAL TIN2009- 14633-C03-01 and FPU AP2007-02867), the Generalitat Valenciana (grant Prometeo/2009/014 and ACOMP/2010/051) and the UPV (grant 20080033).

References

1. M. A. del Agua. Multilingualidad en el reconocimiento de texto manuscrito. Final Degree Project, 2010.
2. D. Ghosh, T. Dube, and P. Shivaprasad. Script Recognition: A Review. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 32(12):2142–2161, Dec. 2010.
3. D. Pérez, L. Tarazón, N. Serrano, F. Castro, O. Ramos-Terrades, and A. Juan. The GERMANA database. In *Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009)*, pages 301–305, Barcelona (Spain).
4. T. Plötz and G. Fink. Markov models for offline handwriting recognition: a survey. *Int. J. on Document Analysis and Recognition (IJDAR)*, 12(4):269–298, Nov. 2009.
5. N. Serrano, A. Giménez, A. Sanchis, and A. Juan. Active learning strategies in handwritten text recognition. In *Proc. of the 12th Int. Conf. on Multimodal Interfaces and the 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010)*, number 86, Beijing (China), Nov. 2010.
6. N. Serrano, D. Pérez, A. Sanchis, and A. Juan. Adaptation from Partially Supervised Handwritten Text Transcriptions. In *Proc. of the 11th Int. Conf. on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2009)*, pages 289–292, Cambridge, MA (USA).
7. N. Serrano, A. Sanchis, and A. Juan. Balancing error and supervision effort in interactive-predictive handwriting recognition. In *Proc. of the Int. Conf. on Intelligent User Interfaces (IUI 2010)*, pages 373–376, Hong Kong (China).
8. N. Serrano, L. Tarazón, D. Pérez, O. Ramos-Terrades, and A. Juan. The GIDOC prototype. In *Proc. of the 10th Int. Workshop on Pattern Recognition in Information Systems (PRIS 2010)*, pages 82–89, Funchal (Portugal).