Character-based Handwritten Text Recognition of Multilingual Documents

Miguel A. del Agua, Nicolás Serrano, Jorge Civera and Alfons Juan

DSIC/ITI, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
{mdelagua,nserrano,jcivera,ajuan}@dsic.upv.es

Abstract. An effective approach to transcribe handwritten text documents is to follow a sequential interactive approach. During the supervision phase, user corrections are incorporated into the system through an ongoing retraining process. In the case of multilingual documents with a high percentage of out-of-vocabulary (OOV) words, two principal issues arise. On the one hand, a minor yet important matter for this interactive approach is to identify the language of the current text line image to be transcribed, as a language dependent recognisers typically performs better than a monolingual recogniser. On the other hand, word-based language models suffer from data scarcity in the presence of a large number of OOV words, degrading their estimation and affecting the performance of the transcription system. In this paper, we successfully tackle both issues deploying character-based language models combined with language identification techniques on an entire 764-page multilingual document. The results obtained significantly reduce previously reported results in terms of transcription error on the same task, but showed that a language dependent approach is not effective on top of character-based recognition of similar languages.

1 Introduction

Have not been until recently when large volumes of old handwritten documents have undergone an image digitalisation process in order to give general public access to this new source of information. However, digitalised handwritten documents cannot be fully exploited by natural language processing (NLP) tools, if texts are not available in electronic format. For this reason, a continuous time-consuming transcription effort is nowadays being carried out by digital libraries.

To alleviate this effort, automatic handwriting transcription techniques based on speech recognition technology have flourished over the last years, although the quality of the transcriptions provided by these techniques is still far from not being in need of supervision [1]. An effective approach to supervision is to integrate an ongoing retraining system that interactively incorporates user corrections once a line has been reviewed. Such a system, along with layout analysis and line detection features, has been implemented in an open source tool called *Gimp-based Interactive transcription of old text DOCuments* (GIDOC) [2].

GIDOC has been used as a platform to develop techniques aimed at reducing user effort and maximise its usability. These techniques range from adapting models from partially supervised transcriptions [3], over an adequate trade-off between error and supervision effort [4], to a variety of active learning strategies to improve the interaction with the user on each new system hypothesis [5].

A specially appealing case in automatic handwritten text recognition is the transcription of multilingual documents. A good example of multilingual document is the GERMANA database [6]. GERMANA is the result of digitizing and annotating a 764-page, single-author manuscript from 1891, written in Spanish up to page 180, but then also written in five other languages, mainly in Catalan and Latin. Another distinctive feature of GERMANA is the large number of out-of-vocabulary (OOV) words accentuated by its multilingual nature. This feature has been the main reason for the relatively poor results obtained so far on the GERMANA database [7].

The work presented in this paper targets both characteristic features of the GERMANA database: Multilinguality and OOV words. Multilinguality is captured by language identification models already discussed in [7]. The problem of OOV words is tackled by deploying character-based n-gram language models. As a consequence, the reported results are the best ever achieved on the GERMANA database.

The rest of this paper is structured as follows. Previous work related to multilinguality and character-based language modelling in speech and handwriting recognition is reviewed in Section 2. In Section 3, the probabilistic framework for language identification on a character-based handwriting recognition approach is presented. Section 4 is devoted to empirical results on the whole GERMANA manuscript. Finally, conclusions and future work are discussed in Section 5.

2 Previous work

Multilinguality in handwritten text recognition arises the challenge of taking advantage of language identification in order to interactively adapt the underlying models of the system and to minimise transcription errors. However, conventional (non-interactive) script and language identification are still in its early stage of research [8], and have remained unexplored until very recently [7].

Preliminary results exploiting multilinguality on the GERMANA database proved the benefits of explicitly modelling language identification at the line level in a interactive transcripcion scenario [7]. However, these results are far from allowing an effective interactive transcription. In that work, the supervision effort would be excesively high, and the user might prefer to ignore the automatically generated output and transcribe the manuscript from scratch. An error analysis revealed that most of these errors were due to out-of-vocabulary (OOV) words. In fact, 53% to 71% of the words in the GERMANA database are singletons, words ocurring only once in the lexicon of each language. Another important problem was the scarse resources available for some languages in the GERMANA database, so as to train their corresponding word-based language models.

The treatment of OOV words is an open problem in different areas of NLP. In speech recognition, which is closely related to handwritten text recognition as far as modelisation is concerned, notable efforts has been deployed over the last decades to deal with OOV words. In [9], the original lexicon is extended with words from external resources that are represented as a sequence of characters (graphemes, to be more precise) converted into phonemes. In [10], several subword based methods for spoken term detection task and phone recognition are presented to search OOV words. Phone and multigram-based systems provide similar performance on the phone recognition task, superseding the standard word-based system.

Regarding handwriting text recognition, the authors in [11] compared the performance of a conventional word-based language model to that of a character-based language model in the context of a German offline handwritten text recognition task. However, character-based language models were not superior to their word-based counterparts. A hybrid approach between a standard character-based n-gram language model and a character-based connectionist language model is proposed in [12], which obtain similar results to word-based systems on the IAM corpus [13].

To the best of our knowledge, character-based language models has not been able so far to supersede word-based language models in handwritten text recognition. Our hypothesis is that tasks tackled in previous work did not contain a significant number of OOV words compared to the figures of the GERMANA database[1]. In GERMANA, the problem of OOV words is aggravated by its multilingual nature, since the presence of languages such as Latin, French, German and Italian is less than 4% of the total number of words. Therefore, the estimation of word-based language models is notably poor, and it is necessary to fall back to adequate character-based language models.

3 Probabilistic framework

Let t be the number of the current text line image to be transcribed, and let x_t be its corresponding sequence of feature vectors. The task of our system is to predict for each text line image first its language label, l_t, and then its transcription, c_t. We assume that all preceding lines have been already annotated in terms of language labels, l_1^{t-1}, and transcriptions, c_1^{t-1}. By application of the Bayes decision rule, the minimum-error system prediction for l_t is:

$$\begin{aligned} l_t^*(x_t, l_1^{t-1}) &= \underset{\tilde{l}_t}{\operatorname{argmax}}\ p(\tilde{l}_t \mid x_t, l_1^{t-1}) \\ &= \underset{\tilde{l}_t}{\operatorname{argmax}}\ p(\tilde{l}_t \mid l_1^{t-1})\, p(x_t \mid \tilde{l}_t) \end{aligned} \tag{1}$$

where in Eq. (1), it is assumed that x_t is conditionally independent of all preceding language labels, l_1^{t-1}, given the current line language label, $\tilde{l}_t$. For the

[1] For example, the IAM corpus only contains about 7% of OOV words.

term $p(x_t \mid \tilde{l}_t)$, we marginalise over all possible character-based transcriptions for language l_t, that is, $C(\tilde{l}_t)$

$$p(x_t \mid \tilde{l}_t) = \sum_{\tilde{c}_t \in C(\tilde{l}_t)} p(\tilde{c}_t \mid \tilde{l}_t)\, p(x_t \mid \tilde{l}_t, \tilde{c}_t) \tag{2}$$

$$\approx \max_{\tilde{c}_t \in C(\tilde{l}_t)} p(\tilde{c}_t \mid \tilde{l}_t)\, p(x_t \mid \tilde{l}_t, \tilde{c}_t). \tag{3}$$

Eq. (3), the Viterbi (maximum) approximation to the sum in Eq. (2), is applied to only consider the most likely transcription. It must be noted that, this language identification technique is one of the most effectives in Automatic Speech Recognition (ASR) [14].

The decision rule (1) requires a *language identification model* for $p(\tilde{l}_t \mid l_1^{t-1})$ and, for each possible language $\tilde{l}_t$, a $\tilde{l}_t$-dependent *character-based language model* for $p(\tilde{c}_t \mid \tilde{l}_t)$ and a $\tilde{l}_t$-dependent *image model* for $p(x_l \mid \tilde{l}_t, \tilde{c}_t)$.

A series of n-gram language identification models were proposed in [7]. In this work, we applied the best performing models, the unigram model

$$\hat{p}(\tilde{l}_t \mid l_{t-1}) = \frac{N(\tilde{l}_t)}{t-1} \tag{4}$$

and the bigram model

$$\hat{p}(\tilde{l}_t \mid l_{t-1}) = \frac{N(l_{t-1}\, \tilde{l}_t)}{N(l_{t-1})}, \tag{5}$$

both estimated by relative frecuency counts, where $N(\cdot)$ denotes the number of occurrences of a given event in the preceding lines, such as the bigram $l_{t-1}\, \tilde{l}_t$ or the unigram $\tilde{l}_t$. It should be noticed that the bigram model makes use of prior knowledge about the GERMANA database, assuming that consecutive lines are usually written in the same language.

A character-based language model for each language $p(\tilde{c}_t \mid \tilde{l}_t)$ is implemented as a conventional *n-gram language model* [15], but considering characters instead of words. Each $\tilde{l}_t$-dependent language model is trained only from those transcriptions labeled with $\tilde{l}_t$. In the case of character-based n-gram language models, the order of the n-gram is normally higher than that employed in word-based models. The aim is to capture information not only regarding intra-word character sequence, but also inter-word relationship, and word tokenisation and segmentation. This information is specially useful in the transcription of OOV words.

Image models for the different languages are implemented in terms of *character HMMs* [2]. Taking advantage that only a single script is used for all the languages considered in the GERMANA database (e.g. Latin), a unique, shared image model is estimated.

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. (3) can be rewritten as

$$l_t^*(x_t, l_1^{t-1}) \approx \operatorname*{argmax}_{\tilde{l}_t} p(\tilde{l}_t \mid l_1^{t-1})^{\beta} \max_{\tilde{c}_l \in C(\tilde{l}_t)} p(x_t \mid \tilde{l}_t)^{\alpha_{\tilde{l}_t}} \tag{6}$$

being

$$p(x_t \mid \tilde{l}_t)^{\alpha_{\tilde{l}_t}} = p(\tilde{c}_t \mid \tilde{l}_t)^{\alpha_{\tilde{l}_t}}\, p(x_t \mid \tilde{l}_t, \tilde{c}_t) \tag{7}$$

where we have introduced an *Identification Scale Factor (ISF)* β and, for each language $\tilde{l}_t$, a language-dependent *Grammar Scale Factor (GSF)* $\alpha_{\tilde{l}_t}$. In the experiments reported below, these parameters are tuned on a validation set.

4 Experiments

Experiments were performed in the GERMANA database [6]. GERMANA is a single-author manuscript from 1891, which contains 764 pages written in up to six different languages. Our main objective is to study the use of character-based models in an interactive transcription task. As it has been said, the utilization of character-based models is motivated by two main features of GERMANA: the high number of OOVs, and the resource scarcity to train robust word language models. In addition, we analyze the performance of the language identification techniques presented in previous section.

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, such as IAM [13]. Also note that the sum of individual lexicon sizes (29.9K) is larger than the size of the global lexicon (27.1K). This is due to presence of words common to different languages, such as 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 an optimised language model on a 10-fold cross-validation experiment.

Table 1. Basic statistics of GERMANA.

Language	Lines	Running Chars	Lexicon	Perplexity
All	20151	1.08M	121	13.1 ± 0.61
Spanish	80.9%	81.2%	114	12.24± 0.15
Catalan	11.8%	11.7%	93	10.39± 0.34
Latin	4.6%	5.2%	91	10.44± 0.36
French	1.3%	1.3%	79	10.96± 0.81
German	1.1%	0.4%	61	10.17± 0.20
Italian	0.3%	0.3%	61	9.44± 0.24

In our experiments, we followed an interactive transcription framework, where the user supervises the output of a system, which is continuosly retrained. To

this purpose, we divided GERMANA in blocks of 500 lines, numbered from 1 to 40. First, blocks number 1 and 2 were manually transcribed and used to build an initial system and tune the training and recognition parameters. Training parameters, such as number of mixture components and states per HMM, remains unchanged in all experiments. Then, starting from block number 3 to the last. First, the language of each line is identified (if needed) and its transcriptions is recognised by the corresponding language dependent system. Next, its transcription and language label is supervised. Finally, after a full new block is supervised, the system is re-trained from all supervised blocks and adapted on the last supervised block. It must be noted that, HMMs image modeling is carried out by the RWTH ASR toolkit [16] and language modeling by SRILM toolkit [15]. We performed two different sets of experiments on the described framework. The objective of the first set was to study the performance of the language identification methods proposed. On other hand, the objective of the second set was to study the transcription accuracy of the system when using each different language identification method.

In the first set of experiments, we compared three different approaches for language identification: *CPL* (simply assigns to a given line the language of the previous one), *unigram* (uses Eq. 4) and *bigram* (uses Eq. 5). We performed the interactive transcription of GERMANA using described framework for each of the approaches. Each time a block is recognised, we measured the number of errors committed by the language identification method used. It must be noted that, in this set of experiments, parameters were tuned to minimise the number of language identification errors. Table 2 shows the results in terms of language identification error-rate (IER) for the whole document. We also included the results on the same framework of the word-based approach presented in [7].

Table 2. Language identification results on GERMANA

System	CPL	Unigram	Bigram
Character-based	2.5	14.2	4.0
Word-based		15.9	5.0

From the results in Table 2, it can be observed that CPL achieved the best performance. CPL took fully advantage of document sequentiality and it only committed errors when the language changed from line to line, which only occurs a few times in GERMANA. In both, character and word based systems, the bigram approach tunned its parameters to ignore the language dependent recogniser probability in Eq. 7 and it forces the system to only relay on the language model probability of language labels. In this case, the bigram approach identifies the language only using the bigram probability. However, the bigram approach only adapts its parameters each time a block is supervised, and thus, it fails to identify all lines of a language when it appears the first time in the transcription process. On the other hand, the character-based unigram approach achieved slightly better results than its word-based version.

In the second set of experiments, we compared five different approaches in terms of Word Error Rate (WER) on recognised transcriptions. WER is defined as the ratio between the minimum number of editing operations to convert the recognised words into the reference, and the number of reference words. In the first approach, we built a *monolingual* system, where we assume all lines to belong to the same language. This approach is considered the baseline, as language identification step is not needed and it is the simplest aproximation to the problem. Next, motivated from the results in [7], we also built four different language dependent systems, which differ on which language identification method is used to switch on the proper language dependent recogniser. All the language dependent systems shared the same HMM image models but differ on their language models, which are only trained from the transcriptions of their corresponding languages. These multilingual systems are named as: *supervised* (language label is manually given), *CPL* (copy previous label), *bigram* (using Eq. (5)), and *unigram* (using Eq. (4)). It must be noted that, in this case, all approaches adapted their parameters to optimize the WER on last block. As the unigram and bigram approaches can be optimized for WER or IER, we also compared the results of both optimizations when transcribing, as the transcriptions produces are different. The results are represented in Fig. 1, in terms of WER of the recognized text up to the current line.

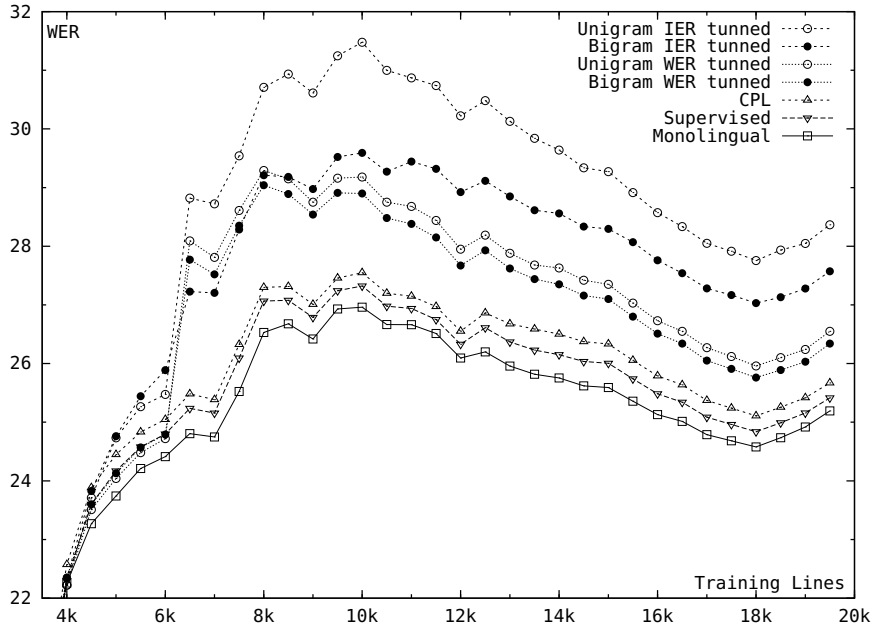

Fig. 1. WER in GERMANA as a function of the number of recognized lines for the monolingual and language-dependent approaches. Results are presented from line 3500, in which a different language apart from Spanish appears.

On the contrary, as it happened in [7], all multilingual systems achieved worse results than the monolingual system. However, even though there is not significant difference between the three best approaches, as corroborated by a bootstrap evaluation [17]; the monolingual approach is considered the best as it is easier to build and it does not need a language identification step in recognition. In error mean terms, even in the supervised approach, where the language is given, the use of language dependent recognizers could not outmatch the monolingual approach. The main cause of the monolingual performance is produced by the origin of all languages but German in GERMANA. Most languages in this document are *Romance* languages, which come from the same original language, sharing a common underlying language structure. For instance, the lexeme of many words can be correctly estimated from the Spanish part in order to recognise other similar romance languages, such as Catalan. In fact, the main responsible of the monolingual result is the high order (9-grams) character-based language model, which was able to estimate the common lexeme structure of all romance languages.

In language dependent approaches, it can be observed that, even though both supervised and CPL approaches achieved the best transcription results, the system performance did not always depend on the language identification performance. On one hand, there is not always a direct relationship between IER and WER. For instance, the unigram and bigram IER optimised approaches achieved a IER of 14.2 and 4.0, respectively, while the WER results were 28.36 and 27.57. On the other hand, as observed from the difference between the different optimizations of unigram and bigram approaches, a system with a worse IER can obtain a better WER results. For example, the bigram WER optimised approach obtained 26.34 of WER from a IER of 8.5, while optimising the IER on the same approach achieved 27.57 of WER from a IER of 4. These results corroborate our conclusions in [7], in which we observed that a language is better recognised using a different language dependent recogniser. However, as said, the monolingual approach achieved better recognition results because the improvement from better estimated languages is already included in the character-based language model.

In terms of transcription performance, in our previous work [7], we also dealt with the complete transcription of GERMANA, but using word-based models. In that case, the monolingual approach obtained 44.39% of WER, however, in this work the same approach obtains 25.19%. These improvement is caused by two factors. On one hand, the RWTH recogniser improved the results due to a new feature extraction method. On the other hand, further error analysis revealed that, as expected, most of this improvement is due to the correct recognition of OOVs words, and punctuation signs. In Figure 2, we can observe the performance of both models in the recognition of a line, concretely, in this example, word-based errors ("estado", "Viuda", and "reflejasen") occured due to OOVs words ("citado", "Vidal", and "refleja"). On the other hand, punctuation signs ("," after "Vidal" and "Reina"), are successfully recognized in the character-based approach, whereas, the word-based approach failed to recognize this signs due

to its scarcity in the training dataset. In past works [6], we only dealt with GERMANA first part, where we reported a performance of 34.51% of WER, in this same partition, the character-based system obtained a performance of 12.12% WER.

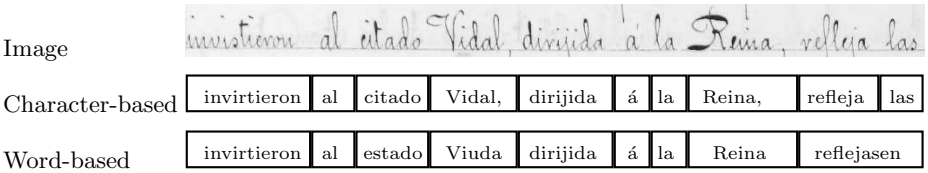

Fig. 2. Comparison of word-based and character-based recognition.

5 Conclusions and future work

We have proposed a character-based approach for interactive transcription of multilingual documents. This approach is motivated by the high number of OOV words in these handwritten text documents. In addition, we have adapted our previous probabilistic framework for language identification in interactive transcription of multilingual documents to be use in a character-based system. Empirical results are presented on the whole GERMANA database, a 764-page, single-author manuscript from 1891 written in up to six different languages. Two different sets of experiments were performed: language identification and automatic recognition experiments. According to the empirical results, in terms of language identification, the simplest technique, that is, the "copy the preceding label" (CPL) bigram model is also the most accurate. On the other hand, in terms of transcription performance, the monolingual approach achieved the best results. This is mainly caused by the use of character-based language models, which successfully estimates the underlying structure of similar languages. We also observed that language identification results did not always correlate with transcription results, and that the use of a language dependent recogniser was not needed in the transcription task proposed. However, a language dependent approach can be useful when dealing with very different languages, which structure do not share any similarities. In addition, the monolingual language model was build from the concatenation of all transcription. A more adequate approach would be to create a mixture of language dependent models, which could improve the monolingual results. Transcription of other multilingual documents remains as future work to better generalise the effectiveness of the presented approach.

Acknowledgements The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287755. Also supported by the Spanish Government (MIPRCV "Consolider Ingenio 2010", iTrans2 TIN2009-14511, MITTRAL TIN2009-14633-C03-01 and FPU AP2007-0286) and the Generalitat Valenciana (Prometeo/2009/014).

References

1. Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(5) (2009) 855 –868
2. Serrano, N., Tarazón, L., Pérez, D., Ramos-Terrades, O., Juan, A.: The GIDOC prototype. In: Proc. of the 10th Int. Workshop on Pattern Recognition in Information Systems (PRIS 2010), Funchal (Portugal) 82–89
3. Serrano, N., Pérez, D., Sanchis, A., Juan, A.: 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), Cambridge, MA (USA) 289–292
4. Serrano, N., Sanchis, A., Juan, A.: Balancing error and supervision effort in interactive-predictive handwriting recognition. In: Proc. of the Int. Conf. on Intelligent User Interfaces (IUI 2010), Hong Kong (China) 373–376
5. Serrano, N., Giménez, A., Sanchis, A., Juan, A.: 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) (November 2010)
6. Pérez, D., Tarazón, L., Serrano, N., Castro, F., Ramos-Terrades, O., Juan., A.: The GERMANA database. In: Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009), Barcelona (Spain) 301–305
7. del Agua, M.A., Serrano, N., Juan, A.: Language identification for interactive handwriting transcription of multilingual documents. In: Proc. of the 5th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2011), Las Palmas de Gran Canaria (Spain) (jun 2011) 596–603
8. Ghosh, D., Dube, T., Shivaprasad, P.: Script Recognition: A Review. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) **32**(12) (December 2010) 2142–2161
9. Bisani, M., Ney, H.: Open vocabulary speech recognition with flat hybrid models. In: Proc. of the European Conf. on Speech Communication and Technology. (2005) 725–728
10. Szoke, I., Burget, L., Cernocky, J., Fapso, M.: Sub-word modeling of out of vocabulary words in spoken term detection. In: Spoken Language Technology Workshop, 2008. SLT 2008. IEEE. (December 2008) 273 –276
11. Brakensiek, A., Rottl, J., Kosmala, A., Rigoll, G.: Off-Line handwriting recognition using various hybrid modeling techniques and character N-Grams. In: In 7th International Workshop on Frontiers in Handwritten Recognition. (2000) 343–352
12. Zamora, F., Castro, M.J., España, S., Gorbe, J.: Unconstrained offline handwriting recognition using connectionist character n-grams. In: Neural Networks (IJCNN), The 2010 International Joint Conference on. (July 2010) 1 –7
13. Marti, U.V., Bunke, H.: The IAM-database: an English sentence database for off-line handwriting recognition. IJDAR (2002) 39–46
14. Schultz, T., Kirchhoff, K.: Multilingual Speech Processing. (2006)
15. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proc. of ICSLP'02. (September 2002) 901–904
16. Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., Ney, H.: The RWTH aachen university open source speech recognition system. In: Interspeech, Brighton, U.K. (September 2009) 2111–2114
17. Efron, B., Tibshirani, R.J.: An Introduction to Bootstrap. Chapman & Hall/CRC (1994)