Handwriting Recognition in Historical Documents using Very Large Vocabularies

Volkmar Frinken
Centro de Visio per Computador
Universidad Autonoma de Barcelona
Edifici O, 08193 Bellaterra, Barcelona, Spain
vfrinken@cvc.uab.es

Andreas Fischer
Centre for Pattern Recognition and Machine Intelligence
Concordia University
1455 de Maisonneuve Blvd West, Montreal, Quebec H3G 1M8, Canada
an_fisch@encs.concordia.ca

Carlos D. Martínez Hinarejos
Escuela Técnica Superior de Ingeniería Informática
Universitat Politècnica de València
Camino de Vera, s/n, 46022 Valencia, Spain
cmartine@dsic.upv.es

ABSTRACT

lets see, to be written *by Volkmar*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque tincidunt fringilla urna, eu congue felis suscipit tincidunt. Aenean semper mi lacus, in laoreet arcu. Nulla velit metus, sodales id tincidunt id, condimentum a neque. Nunc adipiscing nulla egestas risus egestas a convallis leo egestas. Donec vitae libero et risus venenatis vulputate in id magna. Aliquam vel suscipit dolor. Praesent in elit in risus facilisis mattis. Vivamus et urna augue. Duis eu diam magna, sed placerat diam. Aliquam erat volutpat. Mauris pharetra, tellus ut posuere sodales, libero velit tempus odio, nec facilisis urna nisi quis nunc. Sed ut ligula vel turpis hendrerit condimentum. Nunc vel erat a eros tincidunt rutrum eget a massa. Quisque a leo in nibh porttitor fermentum.

Keywords

Historic Documents, Handwriting Recognition, Language Modeling, Google N-grams, BLSTM Neural Networks

1. INTRODUCTION

In the context of preserving the humankind's cultural heritage, large efforts are being done to scan and store vast amount of historical data. The digitization, however, is only the first step on the way to make the contents readily accessible to researchers as well as the general public. A major problem up to date is to extract the textual content from those images into a computer-readable format, which is tedious, time-consuming work and requires expert knowledge.

The last years have seen increased research activities of document analysis for historical data [2, 3] and recent advances have made a (semi-)automatic processing a viable choice for accessing the contents through keyword spotting [5], interactive or full automatic transcription systems [4, 13].

Yet, automatic handwriting recognition is a difficult problem that is not yet solved. Some of the key problems are large varieties in handwriting styles and the need to understand contextual cues through adequate language modeling to resolve ambiguities. Both points are even harder for historic data. Limited amount of transcribed training data for a specific writing style, non-uniform spelling rules, frequent use of abbreviations as well as special symbols can usually be observed. In addition, given an historic text to transcribe, it is very likely that comparable language samples do not exist, as far as time, location, and context is concerned, all of which are important factors when modeling the text via external sources.

In this work we focus on the language modeling aspect and demonstrate a recognition system that uses limited, but accurate n-grams obtained from the training set of the handwriting recognition system and augment the language model with a very large vocabulary obtained from different sources. This maintain the language structure of the training set, which is expected to match the test data, while effectively reducing the out-of-vocabulary rate and significantly increase the recognition rate.

A further contribution of this paper is the presentation of a working recognition system that can cope with very large vocabularies of several hundred thousand words, which is much more than existing system [7, 9], to the knowledge of the authors.

The rest of this article is structured as follows. In Section 2, the database on which we performed the study is introduced. Language modeling and considered corpora are discussed in Section 3 and the handwriting recognition system is ex-

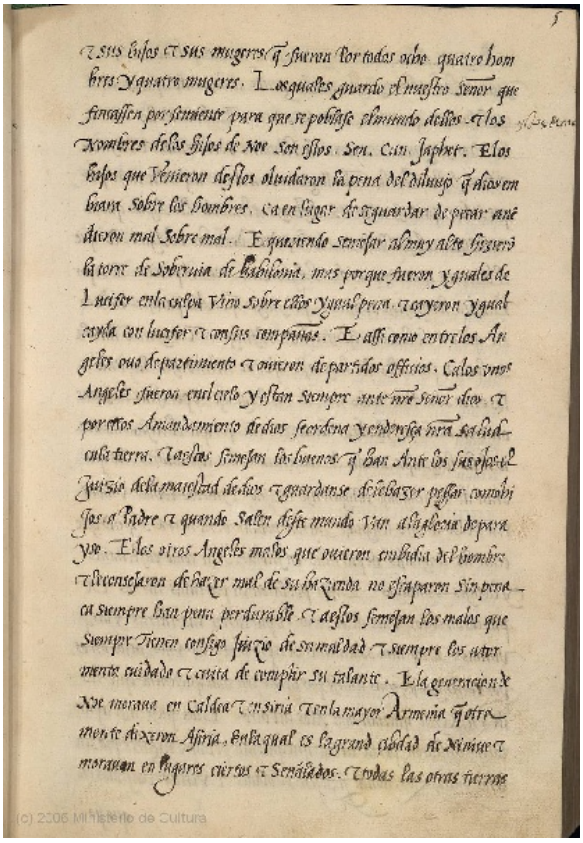

Figure 1: Typical page of the RODRIGO database.

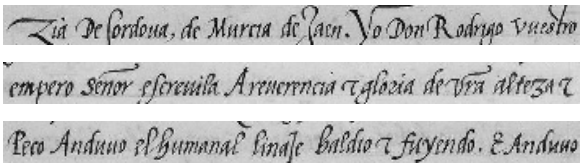

Figure 2: Examples of extracted lines from the RODRIGO database.

plained in Section 4. Section 5 presents an experimental evaluation and Conclusions are drawn in Section 6.

2. DATABASE

The database used in this work is the RODRIGO database [12], which corresponds to a single-writer Spanish text written in 1545, "Historia de España del arçobispo Don Rodrigo". The book has 853 pages with historical chronicles of Spain; most of the pages consist of a single block of well separated lines of calligraphical text. Image in Figure 1 shows an example of a typical page in this manuscript.

Lines in these page were extracted and used as primary data. An example of lines obtained in this extraction process is shown in Figure 2. There is a total of 20356 extracted lines in PNG format files. In the original database they are named after the page number and line number.

The set of lines was divided into three different sets: training (10,000 lines), validation (5010 lines), and test (5346 lines). Test data out-of-vocabulary rate is about 6% with respect to training and validation sets.

3. LANGUAGE MODELING

It has long been known that external information about the target language can help resolve ambiguities and increase the recognition rate [10]. A common choice are statistical bi-gram models that contain a list of words as well as conditional occurrence probabilities $p(w|w')$ of a word w given the previous word w'. With this, the probability of a word sequence $\hat{w} = w_1 \cdots w_N$ can be approximates as

$$p(w_1)p(w_2|w_1)p(w_3|w_2)\cdots p(w_N|w_{N-1}) \quad .$$

This simple, yet powerful model can not grasp long-term dependencies between distance words, but it provides computational advantages, since it fulfills the Markov property.

The actual probabilities are not easily estimated [6]. Just counting the number of observations in a text overestimates rare words that appear by chance while other words that do not appear are assigned a probability of 0. Additionally, a specific text is not a random sampling of words but deals with a certain topic. Hence, any general language model does not reflect the true probabilities and choosing the language corpus is therefore a challenging task. For historical data this problem is even more imminent since words, spelling variants, common abbreviations, and special symbols can change quickly over time and place can lead to a lack of independent data [14].

3.1 Measures

To help in the recognition process, a good language model should assign a high probability to likely sentences. this can be measured in terms of perplexity, which is a function of the average estimated word probability of a given text.

$$\text{ppl}(\hat{w}|LM) = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2\left(p(w_i|w_0\cdots w_{i-1})\right)}$$

where LM is the language model and $\hat{w} = w_1 \cdots w_N$ a word sequence[1]. The lower the perplexity, the higher the average probability and therefore the predictive power of the model.

Note, however, that perplexities can not be easily compared when different underlying vocabularies are used for open vocabulary recognition task, since out-of-dictionary events in the test set are problematic. Usually, those words cannot be recognized, hence their probability would be 0 and the perplexity undefined. Assigning an arbitrary value to OOV words, in turn, does not reflect the transcription process.

Thus, the final recognition rate, given the same underlying recognizer, seems to be a more meaningful measure.

3.2 Google N-Grams

As a byproduct of the massive effort to scan and automatically transcribe millions of printed books, Google has gathered Terra-bytes of textual data and has published n-gram counts for $n = 1 \ldots 5$ for eight of the most wide-spoken languages [11]. Also, the n-gram counts are further subdivided

[1] w_0 is either a token indicating the start of the text or $p(w_1|w_0)$ is defined as $p(w_1)$

into the year of the publication date of the corresponding book.

This renders the data an interesting external source for language information. The text of the database of our experiments (see Section 2) is written in Old Spanish in the 16^{th} century. Unfortunately, only three books from this period are scanned. Also, no manual correction was performed on the data, so that some OCR errors can be observed, i.e. historical long s (s) is often recorded as 'f'.

Thus, we considered a wider time frame to be relevant to in order to create a large vocabulary and reduce the OOV rate in the recognition. High-order n-grams, however, did not improve the perplexity on the validation set in preliminary experiments, so this idea was not further perused.

We chose the year 1800 as a threshold to get a good trade-off between data quality and quantity. This subset consists of 9 388 out of the 854 649 books from the complete Spanish corpus, respectively a list of 1 361 298 unique words, sorted according to their frequency. Even though this is more manageable, 1.36m words are still too much for most automatic recognition systems.

3.3 Don Quixote

Coincidentally, the famous Spanish masterpiece "Don Quixote" was written merely 60 years after the database and can therefore be expected to have at least a similar vocabulary.

An electronic edition of "Don Quijote" by Miguel de Cervantes Saavedra [1] that preserved the original spelling variants [2] and guaranteed to be free of OCR errors served therefore as a second language source. This edition contains 16 538 unique words all of which are added to the vocabulary.

4. BLSTM HANDWRITTEN TEXT RECOGNITION

The recognizer used in this work is based on bidirectional long-short term memory memory neural network [7]. The long-short term memory is a second-order recurrent neural network architecture, in which certain weights of the networks are given by the output of dedicated nodes. By controlling the input, output, and recurrence, a differential version of a memory cell can be simulated. This allows the network to access informations across several time-steps to cope with non-Markovian dependencies. The bidirectionally assures that context from both sides are considered.

The network classifies a sequence of input features into a sequence of posterior character probabilities which are then transformed into the most likely word sequence given a language model through a token passing algorithm.

[2]The famous introductory sentence "En un lugar de la Mancha de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, . . . ", for example, used to be in the original "En vn lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que viuia vn hidalgo de los de lança en astillero, . . . "

Table 1: The list of the seven language models with external vocabulary and the two reference language models along with OOV rates and perplexities on the testing set. Note that the perplexities are only measured on the known words.

Name	ext. vocabulary	OOV rate	perplexity
Ext_{20}	20k	4.70	192.854
Ext_{50}	50k	4.10	205.984
Ext_{100}	100k	3.59	219.128
Ext_{150}	150k	3.30	227.557
Ext_{200}	200k	3.11	233.606
Ext_{250}	250k	2.94	239.044
Ext_{300}	300k	2.80	243.824
Int_{open}	0	6.15	166.741
Int_{closed}	0	0	257.992

4.1 Preprocessing

... we focus on the line level.

- Preprocessing & Features Bunke&Marti *Andreas*

4.2 Training

Training of the BLSTM NN recognizer is done by iteratively adjusting randomly initialized weights via standard backpropagation through time [8]. The objective function is designed to minimize the negative log likelihood of the ground truth, given the network output [7].

4.3 Recognition

word-prefix tree for large vocabularies

Andreas

5. EXPERIMENTAL EVALUATION

5.1 Setup

On order to evaluate the applicability and effectiveness of using very large vocabularies form external sources in the automatic transcription of historic handwritten text, we performed the following set of experiments. We trained 10 BLSTM neural networks on the training set using standard parameters with a learning rate of 10^{-4} and a momentum of 0.9. The number of LSTM notes in the hidden layer was set to 100. These are standard values that turned out to work well for a variety of handwritten data and were not further validated.

After training we selected the single best network according to the character error rate on the validation set. This network was then used to produced the matrices of output character probabilities for each of the lines in the test set. Keeping these fixed, the final transcriptions for the different language models were computed with the Token Passing algorithm from Section 4.3.

All language models use bi-gram probabilities estimated with modified Kneser-Nay smoothing on the training and validation set and differ only in the list of known word. See Table 1 for a summary. An open vocabulary language model Int_{open}

Table 2: The results

LM	WAR	WAR*
Ext_{20}	84.28	88.44
Ext_{50}	84.68	88.30
Ext_{100}	84.91	88.08
Ext_{150}	85.08	88.00
Ext_{200}	85.17	87.90
Ext_{250}	85.21	87.79
Ext_{300}	85.22	87.67
Int_{open}	82.73	88.15
Int_{closed}	89.75	89.75

and a closed vocabulary language model Int_{closed} with additional words from the testing set are compared to seven language models with an additional external vocabulary between 20k and 300k words.

5.2 Results

As can be seen in Table 1, the external vocabulary reduces the out-of-vocabulary rate from 6.15% to 2.80% for the largest of the tested language models and the effect on the recognition accuracies is shown in Fig. 3. One can see that the performance increases with the size of the external vocabulary, but seems to converge at 85.22% achieved with Ext_{300}. The large increase between Int_{closed} and Ext_{20} shows that even a small set of external words can help the recognition substantially. The differences between the four recognitions using Int_{open} (82.73%), Ext_{20} (84.28%), Ext_{300} (85.22%), and Int_{closed} (89.75%) are all statistically significant according to a Student's T-test with $\alpha = 0.05$.

The more words are added, the lower is the out-of-vocabulary rate and more words can potentially recognized. However, with a larger vocabulary the chance for confusing words also increases. To distinguish between theses to effects, a comparison between the word accuracy rate (WAR) and the normalized word accuracy rate (WAR*) is given in Table 2. The normalized word accuracy rate is defined as WAR/(1-OOVRate) and is the recognition accuracy of the subset of words that would be recognized. The addition of a small external dictionary of 20k words increases the absolute and normalized accuracy rate. When further words are added, the normalized accuracy rate decreases. For 300k added words, the benefit of a decreased out-of-vocabulary rate and the risk of recognizing wrong words balance each other out. Hence, a further increase for even larger dictionaries seems unlikely. Note that the achieved accuracies are the highest ever reported on this database, to the knowledge of the authors [REFERENCE].

The decoding with Ext_{300} was also the limit as far as memory resources on the computer system are concerned. The experiments were conducted of a cluster of Intel®Xeon® CPU E5-2665 0 with a clock speed of 2.40GHz and 8GB memory. The average time it took to decode a text line was 24.90s for the Int_{open} language model, 27.30s for Ext_{20} and 39.78s for Ext_{300}.

6. CONCLUSIONS

In this article we described a system for the automatic transcription of historical documents using very large vocabularies gathered from two external sources, the Google N-gram project and an edition of a large, manually transcribed book of the same epoch.

With the inclusion of external language sources, we could significantly reduce the out-of-vocabulary rates from 6.15% to 2.80% and by doing so increase the recognition rate. The positive influence of a larger vocabulary could be observed up to size of 300k external words.

By ordering the words of the lexicon in form of a prefix-tree and using a token passing algorithm in conjunction with a BLSTM neural network, the decoding time for a text line increased by a factor of less than two even when adding very large vocabulary sizes.

This work shows therefore how the drawback of limited available language data for historical documents can be reduced. Future work involves experiments with even larger vocabularies and more sophisticated language models, to further increase the final recognition rate. Additionally, speed-ups for the token passing will be investigated for an even faster recognition.

7. ACKNOWLEDGMENTS

We thank Alex Graves for kindly providing us with the BLSTM Neural Network source code. This work has been supported by the European project FP7-PEOPLE-2008-IAPP: 230653 the European Research Council's Advanced Grant ERC-2010-AdG 20100407, the Spanish R&D projects TIN2009-14633-C03-03, RYC-2009-05031, TIN2011-24631, TIN2012-37475-C02-02 as well as the Swiss National Science Foundation fellowship project PBBEP2_141453.

8. REFERENCES

[1] *El Ingenioso Hidalgo Don Qvixote de la Mancha*, volume 1. `http://www.cervantesvirtual.com/`, Madrid: Gráficas Reunidas, 1928-1931 edition, 1615.

[2] A. Antonacopoulos and A. C. Downton. Special Issue on the Analysis of Historical Documents. *Int'l Journal of Document Analysis and Recognition (IJDAR)*, 9(2-7):75–77, 2007.

[3] B. Barett, M. S. Brown, R. Manmatha, and J. Gehring, editors. *Int'l Conf. on Historic Imaging and Processing*, New York, NY, USA, 2011. ACM Digital Libary.

[4] A. Fischer, M. Wüthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehauser, and M. Stolz. Automatic Transcription of Handwritten Medieval Documents. In *Virtual Systems and Multimedia*, pages 137–142, 2009.

[5] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke. A Novel Word Spotting Method Based on Recurrent Neural Networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(2):211–224, 2012.

[6] J. T. Goodman. A Bit of Progress in Language Modeling: Extended Version. Technical Report MSR-TR-2001-72, Microsoft Resarch, 2001.

[7] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A Novel Connectionist System for Unconstrained Handwriting Recognition.

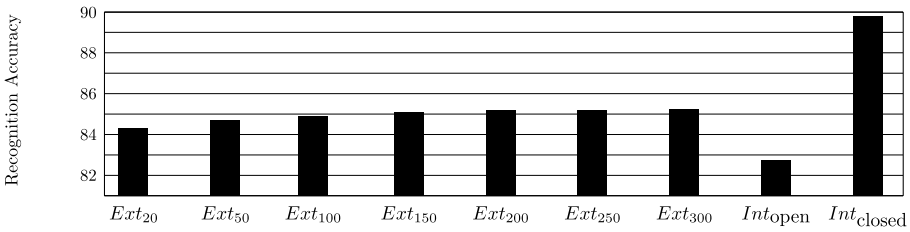

Figure 3: The recognition accuracies using the different language models.

IEEE Transaction on Pattern Analysis and Machine Intelligence, 31(5):855–868, 2009.

[8] A. Graves and J. Schmidhuber. Framewise Phoneme Classification with Bidirectional LSTM Networks. In *Int'l Joint Conf. on Neural Networks*, volume 4, pages 2047–2052, 2005.

[9] M. Kozielski, D. Rybach, S. Hahn, R. Schlüter, and H. Ney. Open Vocabulary Handwriting Recognition Using Combined Word-Level and Character-Level Language Models. In *Int'l Conf. on Acoustics, Speech, and Signal Processing*, page accepted for publication, 2013.

[10] U.-V. Marti and H. Bunke. *Hidden Markov models: Applications in Computer Vision*, chapter Using a Statistical Language Model to Improve the Performance of an HMM-Based Cursive Handwriting Recognition System, pages 65–90. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2002.

[11] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, W. Brockman, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 1 2010.

[12] N. Serrano, F. Castro, and A. Juan. The rodrigo database. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. ELRA.

[13] A. H. Toselli, E. Vidal, and F. Casacuberta. *Multimodal Interactive Pattern Recognition and Applications*. Springer-Verlag, 2011.

[14] M. Wüthrich, M. Liwicki, A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser, and M. Stolz. Language Model Integration for the Recognition of Handwritten Medieval Documents. In *10th Int'l Conf. on Document Analysis and Recognition*, pages 211–215, 2009.