International Conference on Document Analysis and Recognition, Nancy

The LIMSI Handwriting Recognition System for the HTRtS 2014 Contest

Théodore Bluche, Hermann Ney, Christopher Kermorvant

August 25, 2015

UNIVERSITÉ PARIS SUD

Comprendre le monde, construire l'avenir®

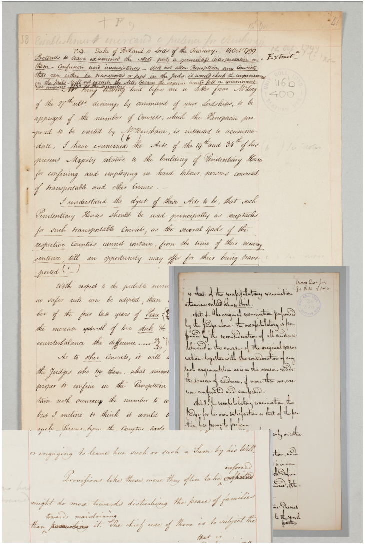
a2ia

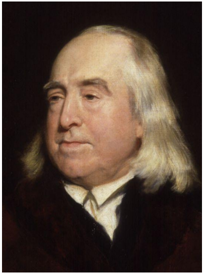
Limsi

The HTRtS 2014 Contest

- **H**andwritten **T**ext **R**ecognition **t**ran**S**criptorium
- Part of the tranScriptorium project aiming at transcribing old manuscripts using HTR systems
- The data comes from the Transcribe Bentham collaborative project

Bentham Manuscripts

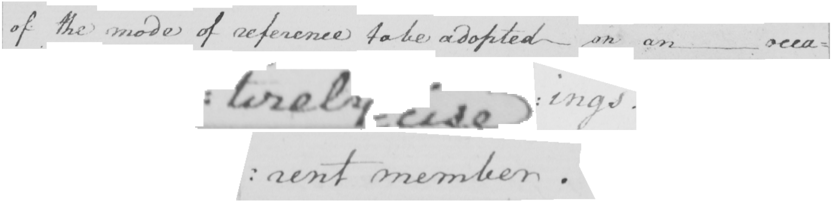




- Manuscripts from J. Bentham (British philosopher, 1748-1832)
- Written by himself and his secretary staff
- About law and moral
- Collected by UCL for the tranScriptorium project

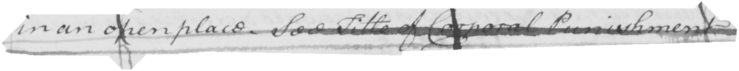
Difficulties

Hyphenations



Crossings

Paper



Overview

The HTRtS 2014 Contest

Data

Set	#Pages	#Lines	#Words	(unique)	#Characters	(unique)
Train	350	9,198	76,707	(12,104)	419,764	(93)
Dev.	50	1,415	11,580		64,070	
Eval.	33	860	7,868		40,231	

- Whole pages are available
- Cropped text lines and their transcript
- Only a few scripters (Bentham + staff)

Evaluation

- Two months to build the systems, one week to produce test set results
- The system performance is measured with the **Word Error Rate** (WER%).

- **Restricted track**: only the provided data are allowed to train the systems
- **Unrestricted track**: participants can use additional data to build the optical and language models

Data Preparation

Image Preprocessing

The systems are trained with text lines cropped from the whole 300 DPI document images.

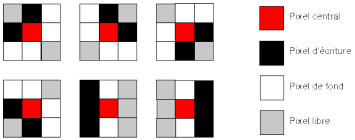
Preprocessing:

- Conversion to gray-level
- Deslant from Buse et al. (1997)
- Contrast enhancement: mapping the 5% darkest pixels to black and 70% lightest ones to white + linear interpolation
- Height normalization to 72px

Feature Extraction

Handcrafted features (Bianne-Bernard, 2011)

- Sliding window of 3px, with 3px step
- **56 handcrafted features** extracted from each frame
  - 8 pixel density measures
  - 12 pixel configurations
  - HoG in 8 directions
  - + deltas (= 28 + 28)



(Bianne-Bernard, 2011)

Pixel values

- Sliding window of 57px, with 3px step
- Rescaled to 25 x 32px (keeps aspect-ratio)
- Extraction of the **800 gray-level pixel intensities** per frame

Language Models: Dealing with Hyphenation

Corpus preparation:

- extraction of complete paragraphs of text
- ignore lines with a single word (consider them as simple paragraph)
- reconstruction of whole words from hyphenated ones

Tokenization:

- split sequences of digits / currency symbols
- isolate punctuation symbols

LM estimation:

- generate $n$gram counts
- for words with unigram counts greater than a threshold: generate all possible hyphenations using Pyphen[1]
- add all word beginnings / endings with the different hyphenation symbols to unigrams with count 1.

$\longrightarrow$ 4gram with Witten-Bell smoothing (Witten & Bell, 1991), vocabulary of 32k words (7k words + hyphenations), 5.5% OOV, ppl 101.

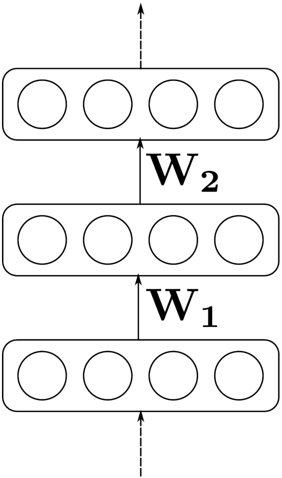
[1] http://pyphen.org/

Decoding

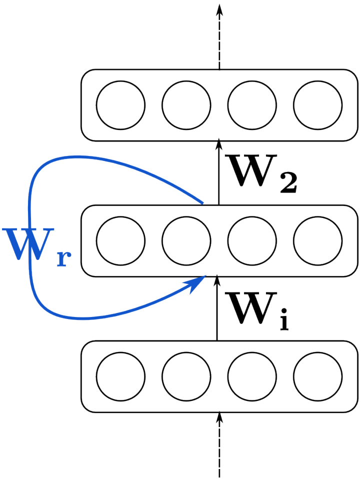
- Hybrid NN/HMMs with $n$gram language models
  - 6-state models for Multi-Layer Perceptrons (framewise cross-entropy training from GMM/HMM alignments)
  - 1-state models for Recurrent Neural Nets (trained with Connectionist Temporal Classification)
- Neural nets predict HMM state q_t
- FST-based decoding with the KALDI Toolkit using scaled posteriors $p(q_t|x_t)/p(q_t)$

Restricted Track: A Combination of Systems

Artificial Neural Networks

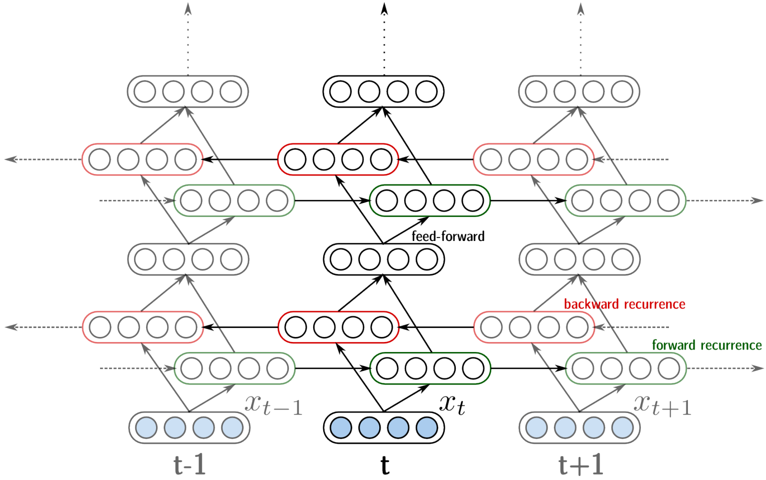


Multi-Layer Perceptron (MLP)



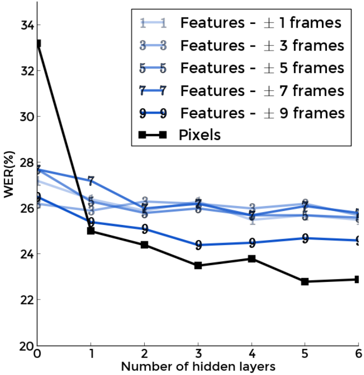
Recurrent Neural Network (RNN)

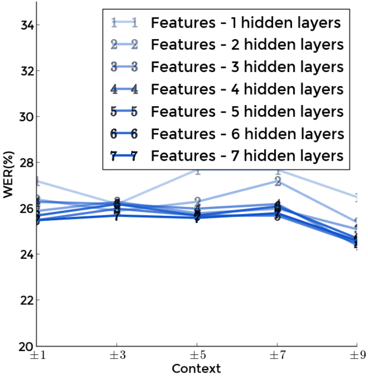
Artificial Neural Networks



Bidirectional RNN

Multi-Layer Perceptrons: impact of depth and context





- 1,024 sigmoid units per layer
- RBM layerwise pre-training with contrastive divergence
- Cross-entropy framewise training from GMM/HMM alignments

Multi-Layer Perceptrons: sequence-training

State-Level Minimum Bayes Risk (sMBR; Kingsbury (2009)), maximize:

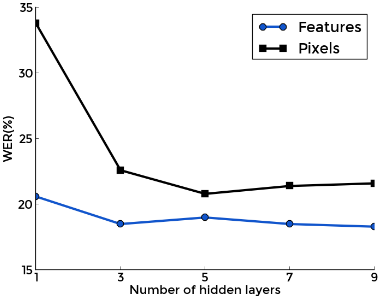
$$E_{sMBR} = \sum_{(\mathbf{x},\mathbf{W}_{ref})\in\mathcal{S}} \frac{\sum_{\mathbf{W}} p(\mathbf{x}|\mathbf{W})P(\mathbf{W})A(\mathbf{W},\mathbf{W}_{ref})}{\sum_{\mathbf{W}'} p(\mathbf{x}|\mathbf{W}')P(\mathbf{W}')}$$

Improvement brought by sMBR sequence training on the validation set

Features	**WER**	**CER**
Handcrafted	21.0%	8.9%
+ sMBR training	**19.4% (-7.6%)**	**7.9% (-11.2%)**
Pixels	22.6%	10.7%
+ sMBR training	**19.9% (-11.9%)**	**8.2% (-23.4%)**

BLSTM-RNNs: impact of depth

Impact of depth (100 units in each layer)



Adding dropout

	Features WER	Features CER
7x100	18.5%	7.5%
7x200	18.0%	7.0%
+ dropout	**17.2%**	**6.7%**
	Pixels	
7x100	21.4%	8.8%
7x200	20.6%	8.4%
+ dropout	**18.7%**	**7.3%**

- 100 tanh units per layer (in each LSTM direction, and each feed-forward)
- no pre-training
- CTC training from character sequences

ROVER / Lattice Combination

Summary of results of restricted systems.

System		WER%	CER%
GMM-HMM	Features	27.9	14.5
Deep MLP	Features	19.4	7.9
	Pixels	19.9	8.2
Deep RNN	Features	17.2	6.7
	Pixels	18.7	7.3

Comparison of combination techniques for the four restricted track systems.

Method	WER%	CER%
ROVER combination (Fiscus, 1997)	16.0	6.6
Lattice combination (Xu et al., 2011)	**15.4**	**5.9**

Restricted track Results

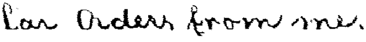
Competition Results for the Restricted Track.

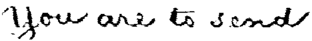
Model		WER%
Deep MLP	Features	19.0
	Pixels	20.0
Deep RNN	Features	17.1
	Pixels	19.0
Lattice combination		15.0
CITlab		**14.6**

Unrestricted Track: A Study of the Importance of Data

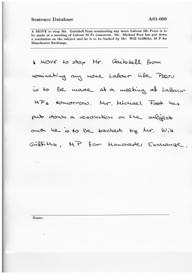
Adding Data to the Training of Optical Models

Georges Washington (`washingtondb-v1.0`)





IAM



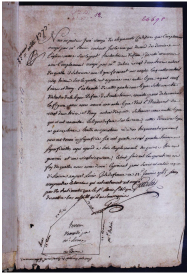
Sentence Database A01-000

A MOVE to stop Mr. Gaitskell from nominating any more Labour life Peers is to be made at a meeting of Labour M Ps tomorrow. Mr. Michael Foot has put down a resolution on the subject and he is to be backed by Mr. Will Griffiths, M P for Manchester Exchange.

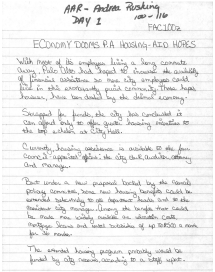
A MOVE to stop Mr. Gaitskell from nominating any more Labour life Peers is to be made at a meeting of Labour MPs tomorrow. Mr. Michael Foot has put down a resolution on the subject and he is to be backed by Mr. Will Griffiths, MP for Manchester Exchange.

Name:

NUMEN



IBM UB 1



AAR - Andrea Rusking
DAY 1 100 - 116
FAC100z

ECONOMY DOOMS P.A HOUSING-AID HOPES

With most of its employees living a long commute away, Palo Alto had hoped to increase the availability of financial assistance so more city employees could live in this exorbitantly priced community. These hopes, however, have been dashed by the dismal economy.

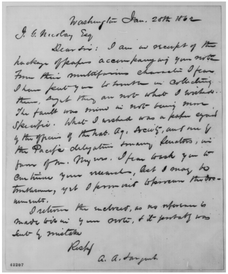
Strapped for funds, the city has concluded it can afford only to offer greater housing incentives to the top echelon at City Hall.

Currently, housing assistance is available to the four council-appointed offices: the city clerk, auditor, attorney and manager.

But under a new proposal backed by the council's policy committee, some new housing benefits could be extended selectively to all department heads and to the assistant city manager. Among the benefits that could be made more widely available are relocation costs, mortgage loans and rental subsidies of up to $500 a month for 36 months.

The extended housing program probably would be funded by city reserves, according to a staff report.

Abraham Lincoln



Adding Data to the Training of Optical Models

Collecting Data

Data used for optical model training.

Track	Name	Number of text lines
Restricted	Bentham	9,198
Unrestricted	IAM	6,482
	NUMEN	11,710
	G. Washington (GW)	642
	IBM UB 1	825
	A. Lincoln (AL)	3,960

Generating Annotations

For IBM and AL, the line positions are unknown, we only have the images and the transcripts

$\longrightarrow$ automatic line segmentation and ground-truth alignment using the technique presented in (Bluche et al., 2014)

Effect of Adding Data to the Training of Optical Models

- **uRNN1** : Bentham, G. Washington, subset of IAM, Numen and A. Lincoln
- **uRNN2** : Bentham, G. Washington, subset of IAM, Numen, IBM and A. Lincoln
- **uRNN3** : Bentham, G. Washington, IAM, IBM, A. Lincoln and Numen

Name	Training data	RNN-CER%	WER%
RNN features	Bentham	8.9	17.2
uRNN1	Bentham, GW, sIAM, sNumen, sAL	7.5	16.5
uRNN2	Bentham, GW, sIAM, sNumen, sIBM, sAL	6.6	15.8
uRNN3	Bentham, GW, IAM, Numen, IBM, AL	6.6	15.8

Effect of Adding Data to the Training of Language Models

- Adding the Open American National Corpus (OANC, Ide & Suderman (2007))
- Vocabulary: 110k words (80k words + hyphenations), 2.5% OOV
- 2gram LM and lattice rescoring with 3grams, ppl 250

Improvements brought by adding more LM data (WER% / CER%;).

		Restricted LM	Unrestricted LM
Deep MLP	Features	19.4 / 7.9	16.7 / 6.9
	Pixels	19.9 / 8.2	17.5 / 7.5
Deep RNN	Features	17.2 / 6.7	14.9 / 5.7
	Pixels	18.7 / 7.3	16.3 / 6.4
Lattice combination		15.4 / 5.9	12.5 / 4.9
	uRNN1	16.5 / 6.1	13.4 / 5.1
	uRNN2	15.8 / 5.6	13.1 / 4.8
	uRNN3	15.8 / 5.6	13.1 / 4.8
Lattice combination		14.6 / 5.4	11.8 / 4.8

Results of the Unrestricted Track

Competition Results for the Unrestricted Track.

Model	WER%
RNN features	14.7
uRNN1	12.9
uRNN2	12.7
uRNN3	12.4
Lattice Combination	11.1
A2iA production system	**8.6**

Post-Evaluation Improvements, and the HTRtS 2015 Contest

A More "Author-Specific" Language Model

- The A2iA production system used Bentham texts retrieved from the web
- OOV rate 1.5%, ppl 215

WER% improvements brought by adding even more LM data.

		OANC	Bentham texts
Deep MLP	Features	16.7	14.0
	Pixels	17.5	14.6
Deep RNN	Features	14.9	13.1
	Pixels	16.3	14.4
Lattice combination		12.5	10.7
	uRNN1	13.4	11.9
	uRNN2	13.1	11.3
	uRNN3	13.1	11.3
Lattice combination		11.8	9.7

A More Careful Training of Neural Networks

After the evaluation, more time to tune hyper-parameters

- MLP: choice of units per layer, size of context, etc.
- RNN: size of hidden layers, better use of dropout (cf Bluche et al. (2015))

WER% of the refined models (restricted track, validation set).

		Competition	Refined
Deep MLP	Features	19.4	18.6
	Pixels	19.9	19.2
Deep RNN	Features	17.2	16.2
	Pixels	18.7	16.9
Lattice combination		15.4	14.6

Overview of the System for HTRtS 2015

New data

- The validation set became part of the training set, the evaluation set became validation set
- Doubled amount of training data **but without line positions** in the page (using (Bluche et al., 2014) to segment/align)
- Features from Kozielski et al. (2013) provided by organizers with the data

New system

- We built the same RNN architecture as 2014 + subsampling
- Sliding window of 4px and shift 2px for all features (handcrafted, pixels, and provided)
- We trained one RNN for each feature set
- **early combination**: remove the top layer of each RNN and add a shared LSTM layer one top of all three RNNs
- We built a hybrid word/character LM (word trigram with 5k/15k vocab., char 7gram)

Post-Evaluation Results

Restricted track

Model		WER%	CER%
MLP	Features	18.6	7.5
	Pixels	20.9	8.2
RNN	Features	16.2	5.4
	Pixels	16.9	5.9
Lattice combination		**14.1**	**5.0**
CITlab		14.6	-
Ours (Competition)		15.1	-
HTRtS 2015 system*		**8.7**	**2.8**

Unrestricted track

Model		WER%	CER%
MLP	Features	13.2	4.9
	Pixels	14.4	6.1
RNN	Features	11.2	4.0
	Pixels	11.5	4.4
	uRNN1	10.9	4.0
	uRNN2	10.5	3.7
	uRNN3	10.2	3.6
Lattice combination		**8.6**	**3.1**
A2iA prod.		**8.6**	-
Ours (Competition)		11.1	-
HTRtS 2015 system*		**7.6**	**2.6**

* : more training data, and 2014 evaluation set was the validation set for 2015

Conclusion

Conclusion

- We were the only team to submit results in both tracks and ranked second in each
- The contest was a good opportunity to
  - compare two kinds of inputs (features and pixels)
  - compare two kinds of NN optical models (MLP and RNNs)
  - try different combination methods
  - study the impact of added training data for optical and language models
- We also proposed a simple way of dealing with hyphenation

Thank you for your attention!

tb@a2ia.com

References

Bianne-Bernard, A.-L. (2011). Reconnaissance de mots manuscrits cursifs par modèles de Markov cachés en contexte. Ph.D. thesis, Telecom ParisTech.

Bluche, T., Kermorvant, C., & Louradour, J. (2015). Where to Apply Dropout in Recurrent Neural Networks for Handwriting Recognition? In 13th International Conference on Document Analysis and Recognition (ICDAR), (pp. --). IEEE.

Bluche, T., Moysset, B., & Kermorvant, C. (2014). Automatic Line Segmentation and Ground-Truth Alignment of Handwritten Documents. In 14th International Conference on Frontiers in Handwriting Recognition (ICFHR2014), (pp. 667--672).

Buse, R., Liu, Z. Q., & Caelli, T. (1997). A structural and relational approach to handwritten word recognition. IEEE Transactions on Systems, Man and Cybernetics, 27(5), 847--61.
URL http://www.ncbi.nlm.nih.gov/pubmed/18263093

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on, (pp. 347--354). IEEE.

Ide, N., & Suderman, K. (2007). The open american national corpus (oanc).
URL http://www.americannationalcorpus.org/OANC/index.html

Kingsbury, B. (2009). Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009., (pp. 3761--3764). IEEE.

Kozielski, M., Doetsch, P., & Ney, H. (2013). Improvements in RWTH ' s system for off-line handwriting recognition. In International Conference on Document Analysis and Recognition (ICDAR).

Witten, I. H., & Bell, T. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. Information Theory, IEEE Transactions on, 37(4), 1085--1094.

Xu, H., Povey, D., Mangu, L., & Zhu, J. (2011). Minimum bayes risk decoding and system combination based on a recursion for edit distance. Computer Speech & Language, 25(4), 802--828.

MLP WERs

Word Error Rates of DNNs trained with cross-entropy, with different number of hidden layers and different inputs. on the validation set. The best systems are indicated in bold face.

		Number of hidden layers						
Features	**Context**	1	2	3	4	5	6	7
Hand-crafted	± 1	27.2	26.4	25.9	26.3	25.5	25.7	25.5
	± 3	26.2	25.9	26.3	26.2	26.0	26.2	25.7
	± 5	27.7	26.3	25.8	26.0	25.7	25.7	25.6
	± 7	27.7	27.2	26.0	26.2	25.7	26.1	25.8
	± 9	26.5	25.4	25.1	**24.4**	24.5	24.7	24.6
Pixels	-	33.2	25.0	24.4	23.5	23.8	**22.8**	22.9

RNN WERs

RNNs on handcrafted and pixel features (results on the validation set, R-CER is the CER of the RNN alone, without LM).

	Handcrafted Features			**Pixels**		
	R-CER	WER	CER	R-CER	WER	CER
1x100	17.3	20.6	8.8	38.9	33.8	19.6
3x100	12.8	18.5	7.5	17.7	22.6	10.2
5x100	12.0	19.0	7.6	14.0	20.8	8.7
5x200	11.8	19.9	7.7	14.0	21.4	8.9
7x100	11.1	18.5	7.5	12.2	21.4	8.8
7x200	11.0	18.0	7.0	11.8	20.6	8.4
7x200 + dropout	8.9	**17.2**	6.7	9.2	**18.7**	7.3
9x100	10.6	18.3	7.3	12.1	21.6	8.9