

ICDAR–2009 Tutorial:

Interactive Multimodal Transcription of Text Images II – Computer-Assisted Transcription of Text Images (CATTI)

Alejandro H. Toselli & Enrique Vidal

atoselli@irisa.fr (on leave from PRHLT)

evidal@iti.upv.es



Pattern Recognition and Human Language Technology Group

Instituto Tecnológico de Informática – Universidad Politécnica de Valencia



Spain

July 2009

ICDAR09: Interactive Multimodal Transcription

A B L A N K P A G E

Tutorial Contents and Schedule

- I Introduction
 - Multimodal Interaction in Pattern Recognition
 - Interactive-Predictive Pattern Recognition and Document Image Analysis
 - Quick Survey of Handwritten Text Recognition (HTR) concepts and techniques
- I-p Off-line HTR in practice
 - HTR Preprocessing
 - Training HMMs using the "Hidden Markov Model Toolkit" (HTK)
 - Training Language Models and Dictionaries for HTR
 - HTR Experiments
- II **Computer-Assisted Transcription of Text Images (CATTI)**
 - Human interaction in HTR
 - A CATTI formal framework
 - Feedback-derived dynamic language modelling and search
 - Performance measures and results achieved in typical applications
- *** *COFFEE BREAK*
- II-p CATTI in practice
 - Adapting Language Models and Search for CATTI
 - CATTI Experiments
 - Analyzing quantitatively the CATTI performance
- III Multimodality in CATTI (MM-CATTI)
 - Touchscreen based multimodal user correction
 - A MM-CATTI formal framework
 - Multimodal language modelling and search
 - Performance measures and results achieved in typical applications
- III-p Demonstration of a complete MM-CATTI System in a real HTR task

Index

- 1 Human interaction in HTR ▷ 3
- 2 A CATTI formal framework ▷ 6
- 3 Feedback-derived dynamic language modelling and search ▷ 8
- 4 Performance measures and Corpora used in the experiments ▷ 10
- 5 Results ▷ 15
- 6 Bibliography ▷ 17

Computer Assisted Transcription of Text Images (CATTI)

- Current cursive (off-line) HTR systems are far from being perfect and generally need human post-editing to check and correct the results of such systems
- In a computer-assisted, interactive framework, rather than full automation, the system aims to facilitate and speed up the human transcription task
- This framework, called “CATTI”, combines the efficiency of automatic handwriting recognition systems with the accuracy of the experts, leading to a cost-effective perfect transcription results
- CATTI can be properly formulated within the general *Interactive-Predictive Pattern Recognition* paradigm
- Performance measures for CATTI should aim at estimating human-effort, rather than error rate


A B L A N K P A G E

How does CATTI work?

- The HTR system proposes a full transcription of the input handwritten text line image.
- The human transcriber (user) validates a prefix of the transcription which is error-free.
- The human enters a word (or words) to correct the erroneous text, producing a new prefix.
- The HTR suggests a suitable continuation to this prefix.
- This previous four steps are iterated until a final, perfect transcription is produced.

Every correction made by the user helps the system automatically avoid further errors in the suggested text.

CATTI operation example

	x	
STEP-0	p	
STEP-1	$\hat{s} \equiv \hat{w}$	antiguas ciudadelas que en el Castillo sus llamadas
	p'	antigu
	κ	os
	p	antiguos
STEP-2	\hat{s}	antiguos ciudadanos que en el Castillo sus llamadas
	p'	antiguos ciudadanos que en
	κ	Castilla
	p	antiguos ciudadanos que en Castilla
FINAL	\hat{s}	antiguos ciudadanos que en Castilla se llamaban
	p'	antiguos ciudadanos que en Castilla se llamaban
	κ	
	$p \equiv T$	antiguos ciudadanos que en Castilla se llamaban

Post-editing WER: 6/7 (86%)

Interactive WSR: 2/7 (29%, assuming a whole-word correction in step-1)

Estimated effort reduction: $1 - 29/86$ (66%).

Statistical framework for CATTI

Given a feature vector stream, x , a set of morphological, lexicon and language models, \mathcal{M} and a *transcription prefix*, p , validated by the user in the previous step, obtain a proper completion (*suffix*) of p from which x can be produced with maximum likelihood; that is:

$$\hat{s} = \underset{s}{\operatorname{argmax}} P_{\mathcal{M}}(s | x, p)$$

Using the Bayes theorem (and dropping \mathcal{M} to simplify notation):

$$\hat{s} = \underset{s}{\operatorname{argmax}} P(x | p, s) \cdot P(s | p)$$

In practice, a *Grammar Scale Factor* is generally used:

$$\hat{s} = \underset{s}{\operatorname{argmax}} P(x | p, s)^{(1-\alpha)} \cdot P(s | p)^\alpha$$

Statistical framework for CATTI (cont.)

Following the prefix-suffix assumption, x can be considered split into two fragments, x_1^b and x_{b+1}^m , where m is the length of x .

This allow us to marginalize $Pr(x | p, s)$ on the boundary point, b , leading to:

$$\hat{s} = \underset{s}{\operatorname{argmax}} \sum_{1 \leq b \leq m} \Pr(x, b | p, s) \cdot \Pr(s | p)$$

Now (realistically) assuming that $Pr(x_1^b | p, s)$ does not depend on s and $Pr(x_{b+1}^m | p, s)$ does not depend on p :

$$\hat{s} \approx \underset{s}{\operatorname{argmax}} \sum_{1 \leq b \leq m} \Pr(x_1^b | p) \cdot \Pr(x_{b+1}^m | s) \cdot \Pr(s | p)$$

And approximating the sum by the dominating term:

$$\hat{s} \approx \underset{s}{\operatorname{argmax}} \max_{1 \leq b \leq m} \Pr(x_1^b | p) \cdot \Pr(x_{b+1}^m | s) \cdot \Pr(s | p)$$

- $Pr(x_1^b | p)$, $Pr(x_{b+1}^m | s)$: *conventional morphological word HMMs*
- $Pr(s | p)$: *prefix-conditioned Language Model*

CATTI Models

N-Gram Language Modeling:

Let $w = w_1^l$ be a full sentence hypothesis and $p = w_1^k, s = w_{k+1}^l$.

$$Pr(s | p) = \frac{Pr(s, p)}{Pr(p)} = \frac{Pr(w)}{Pr(p)} \approx \frac{\prod_{i=1}^l Pr(w_i | w_{i-N+1}^{i-1})}{\prod_{j=1}^k Pr(w_j | w_{j-N+1}^{j-1})}$$

$$= \prod_{i=k+1}^l Pr(w_i | w_{i-N+1}^{i-1})$$

The terms from $k + 1$ to $k + N - 1$ include dependences from the already known words w_{k-N+2}^k . The remaining terms are usual N-Grams; that is:

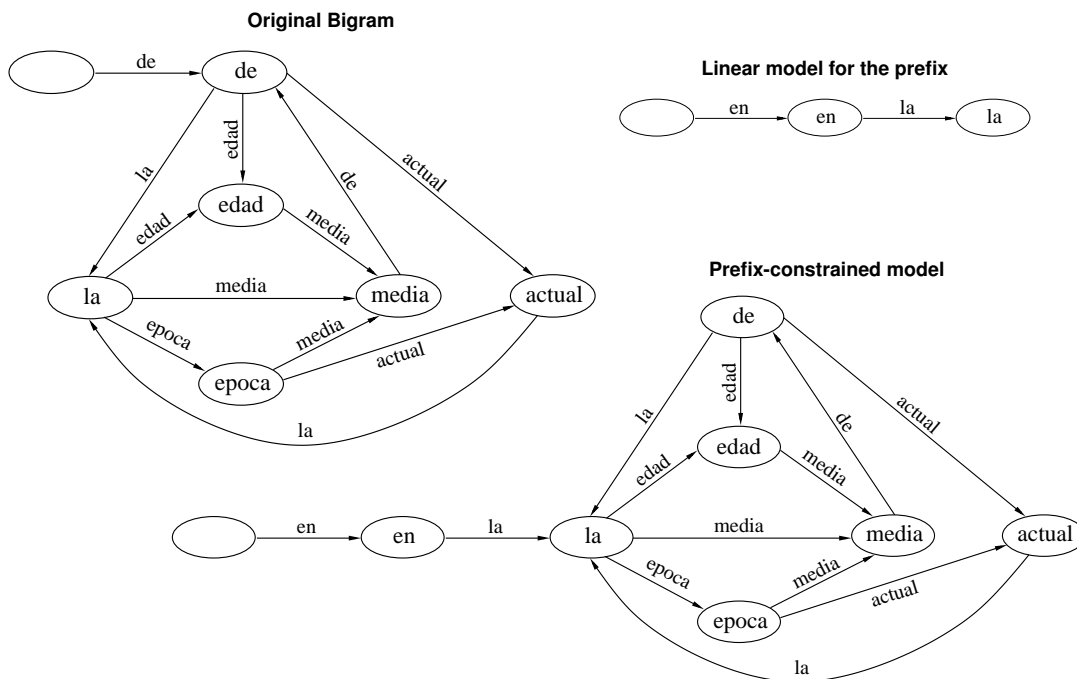
$$P(s | p) \approx \prod_{i=k+1}^{k+N-1} Pr(w_i | w_{i-N+1}^{i-1}) \cdot \prod_{i=k+N}^l Pr(w_i | w_{i-N+1}^{i-1})$$

CATTI “Dynamic” Language Modeling

Training samples

- de la edad media
- de edad media
- de la epoca media
- de la epoca actual
- de la media
- de la actual
- de actual

Prefix = en la



CATTI “dynamic language model” building. A *prefix-constrained model* is obtained by concatenating a *bigram* trained from the training samples to a *linear model* which accounts for the prefix “en la”.

Performance Measures for CATTI

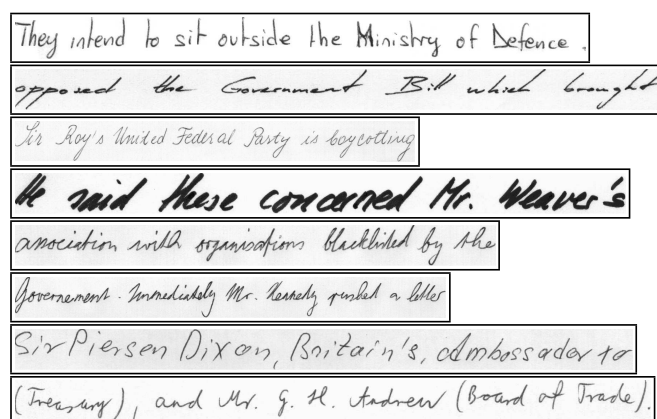
- **WORD ERROR RATE (WER):**
Minimum number of *non-interactive word* corrections (insertions, deletions and substitutions) needed to edit the system output into a (single) target reference
- **WORD STROKE RATIO (WSR):**
Minimum number of word corrections that a (hypothetical) user would have to interactively make to achieve a given reference transcription, divided by the overall number of reference words.
- **KEY STROKE RATIO (KSR):**
Number of characters that, according to a reference transcription, should have to be *interactively* typed by the user, divided by the overall number of reference characters

The relative difference between WSR and WER estimates the human effort that CATTI would save, with respect to that of classical HTR followed by post-editing.

Corpora for HTR experiments: IAMDB

Handwritten texts from the Lancaster-Oslo/Bergen Corpus (LOB)

Publicly available: www.iam.unibe.ch/fki/databases



Number of:	Training	Test	Total	Lexicon	OOV	Tr. Ratio
writers	448	100	548	–	–	–
sentences	2 124	200	2 324	–	–	–
words	42 832	3 957	46 789	8 017	921	128
characters	216 774	20 726	237 500	78	0	2 779

LM training data: approx. 10^6 **running words** from the LOB corpus

Corpora for HTR experiments: ODEC

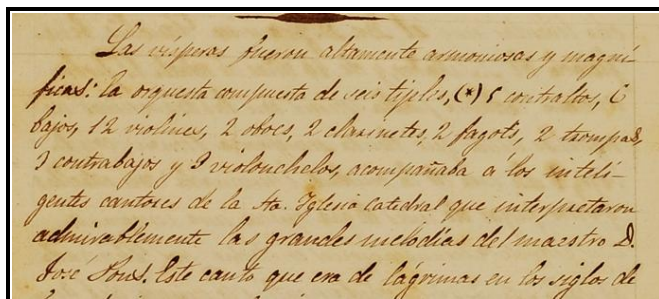
Answers extracted from survey forms made for a telecommunication company

<p>DIFFERENT STYLES <i>r mejorar el servi</i> BEBIAN DAR MÁS PACTI establecimiento de llau concenonio es que es funcionamiento desservic</p>	<p>DIFFICULT LINE SEPARATION DAR LA OPORTUNIDAD A OVE UNA PERSONA OVE LLEVA BASTANTE TIEMPO CONTEMPLANDO EN TELEFONIA OPERABLE OTRO TP. MD</p>
<p>UNUSUAL ABBREVIATIONS TELEF. TEND. REC. CL. EL. SERVIC.</p>	<p>VARIABLE STROKE THICKNESS INFORMACION E ANDO LLAM <i>En mi despacho</i></p>
<p>CROSSED-OUT WORDS EL MONDO DE OVE E AVÉCES L REBASAR EL NÚ</p>	<p>ORTHOGRAPHIC MISTAKES CANVIAR ESCESIVAS HUTILIZAR FALTURA HESTOI VAJAR</p>

Number of:	Training	Test	Total	Lexicon	OOV	Tr. Ratio
writers/sentences	676	237	913	—	—	—
words	12287	4084	16371	2790	518	4.4
characters	64 666	21 533	86 199	80	0	808

Corpora for HTR experiments: “Cristo Salvador”(CS)

Single writer manuscript from the XIX century



Corpora for HTR experiments: CS partitions

CS Page partition

Number of:	Training	Test	Total	Lexicon	OOV	Tr. Ratio
pages	53	53	53	–	–	–
text lines	681	491	1 172	–	–	–
words	6 432	4 479	10 911	2 623	1 313	2.5
characters	36 699	25 460	62 159	78	0	470

CS Book partition

Number of:	Training	Test	Total	Lexicon	OOV	Tr. Ratio
pages	33	20	53	–	–	–
text lines	675	497	1 172	–	–	–
words	6 222	4 689	10 911	2 536	1 400	2.5
characters	35 845	26 314	62 159	78	0	460

Non-interactive HTR baseline results

WER obtained with *closed vocabulary* for different corpora: IAMDB, ODEC and CS (*book* and *page* partitions).

- No case distinction or diacritics; no punctuation marks.
- Character HMMs: 6 states, 64 Gaussian densities per state
- Language models: Bi-grams

Corpus		IAMDB	ODEC	CS-page	CS-book
Writers		many	many	1	1
HMMs	Characters	78	80	78	78
	Tr. Ratio	2 779	808	470	460
Lang. Model	Lexicon	8 017	2 790	2 623	2 536
	OOV	921	518	1 313	1 400
	Tr. Ratio	128	4.4	2.5	2.5
WER (%)		25.8	25.0	34.1	38.8

CATTI Results

In all the experiments, only interaction at the *word level* is assumed; that is, each interaction step involves the correction of a *single, whole word* from the system-predicted suffix.

This allows proper comparisons of the *estimated user effort* needed for non-interactive post-editing (WER) versus interactive processing (WSR).

Performance of the baseline off-line HTR system (WER) and the CATTI system (WSR) for different tasks. 2-gram language models in all the cases:

	IAMDB	ODEC	CS	
			<i>page</i>	<i>book</i>
WER (%)	25.8	25.0	34.1	38.8
WSR (%)	21.8	19.4	32.1	36.8
Effort-Reduction (%)	16	22	6	5

Bibliography

- E. Vidal, F. Casacuberta, L. Rodríguez, J. Civera and C. Martínez. “Computer-assisted translation using speech recognition”. IEEE Trans. on Audio, Speech and Language Processing, 14(3):941-951, 2006.
- E. Vidal, L. Rodríguez, F. Casacuberta and I. García-Varea: “Interactive Pattern Recognition”. 4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI-07), Volume 4892 of LNCS, pp.60-71. Brno, Czech Republic, June 2007.
- L. Rodriguez, F. Casacuberta, and E. Vidal. “Computer Assisted Transcription of Speech” Proc. of the third Iberian Conference on Pattern Recognition and Image Analysis, Volume 4477 of LNCS, pp.241-248, Girona (Spain), June 2007.
- A.H. Toselli, V. Romero, L. Rodríguez and E. Vidal. “Computer Assisted Transcription of Handwritten Text”. 9th Int. Conference on Document Analysis and Recognition (ICDAR 2007), pp.944-948. IEEE Computer Society, Curitiba, Paraná (Brazil), September 2007.
- V. Romero, A.H. Toselli, L. Rodríguez and E. Vidal. “Computer Assisted Transcription for Ancient Text Images”. Int. Conference on Image Analysis and Recognition (ICIAR 2007), volume 4633 of LNCS, pp.1182-1193. Montreal (Canada), August 2007.