ICFHR 2014 Tutorial:

Handwritten Text Recognition: Word-Graphs, Keyword Spotting and Computer Assisted Transcription

V - WG Applications: Computer-Assisted Transcription of Text Images

Moisés Pastor, Joan A. Sánchez, Alejandro H. Toselli and Enrique Vidal

Pattern Recognition and Human Language Technology Research Center
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia

{mpastorg,jandreu,ahector,evidal}@prhlt.upv.es

September 1st, 2014

Outline

Are current HTR results useful?

- Practically useful results for small vocabulary and/or syntax restricted text (e.g., recognition of bank check legal amounts, form-constrained text, etc.)
- Results worsen with large vocabularies and/of unrestricted text (e.g., comments in survey questionaries, historic books, etc.): In practice, word accuracy ranges from 85% to 30%, or even worse for difficult texts, noisy images and/or insufficient training data.
- Such accuracy could be enough for tasks such as document *indexing and searching* in some (or many?) applications.
- But it is too low for *high quality transcription* of most handwritten text images of interest.
  - Human *post-editing can be very expensive* and hardly acceptable by profesional transcribers (paleographers, e.g.).
  - *Computer Assisted, Interactive-Predictive processing* offers promise for *improvements in practical performance and user acceptance*.

Computer Assisted Transcription of Text Images (CATTI)

- In a computer-assisted, interactive-predictive framework, rather than full automation, the system should aim to easy and speed up the human transcription task
- This framework combines the efficiency of automatic handwriting recognition systems with the accuracy of the experts, leading to a cost-effective perfect transcription results
- Rather than measuring error rate, performance evaluation for CATTI should aim at estimating interaction human-effort.

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.

CATTI operation example

	x	
STEP-0	p	
STEP-1	$\hat{s} \equiv \hat{w}$	**antiguas cuidadelas que en el Castillo sus llamadas**
	p'	**antigu**
	κ	**os**
	p	**antiguos**
STEP-2	$\hat{s}$	**ciudadanos que en el Castillo sus llamadas**
	p'	**antiguos ciudadanos que en**
	κ	**Castilla**
	p	**antiguos ciudadanos que en Castilla**
FINAL	$\hat{s}$	**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:

$$\hat{s} \;=\; \arg\max_{s} P_{\mathcal{M}}(s \mid x, p)$$

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

$$\hat{s} \;=\; \arg\max_{s} P(x \mid p, s) \cdot P(s \mid p)$$

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

$$\hat{s} \;=\; \arg\max_{s} P(x \mid p, s)^{(1-\alpha)} \cdot P(s \mid p)^{\alpha}$$

Statistical framework for CATTI (cont.)

HTR main equation:

$$\widehat{w} \;=\; \arg\max_{w} P(x \mid w) \cdot P(w)$$

CATTI main equation:

$$\hat{s} \;=\; \arg\max_{s} P(x \mid p, s) \cdot P(s \mid p)$$

The concatenation of p and s is the whole sentence, w. Therefore CATTI is very similar to the basic HTR. Two main differences:

- The maximization in CATTI must be carried out only over suffixes of p, rather than over whole sentences
- $P(s \mid p)$ in CATTI is interpreted as a kind of of "dynamic" language model, conditioned by increasingly long prefixes, rather than the "static" HTR language model $P(w)$.

CATTI decoding

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 $P(x \mid p, s)$ on the boundary point, b, leading to:

$$\hat{s} = \arg\max_s \sum_{1 \le b \le m} P(x, b \,|\, p, s) \cdot P(s \,|\, p) = \arg\max_s \sum_{1 \le b \le m} P(x_1^b, x_{b+1}^m \,|\, p, s) \cdot P(s \,|\, p)$$

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

$$\hat{s} \approx \arg\max_s \sum_{1 \le b \le m} P(x_1^b \,|\, p) \cdot P(x_{b+1}^m \,|\, s) \cdot P(s \,|\, p)$$

And approximating the sum by the dominating term:

$$\hat{s} \approx \arg\max_s \max_{1 \le b \le m} P(x_1^b \,|\, p) \cdot P(x_{b+1}^m \,|\, s) \cdot P(s \,|\, p)$$

CATTI Models

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

N-Gram Language Modeling:

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

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

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

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:

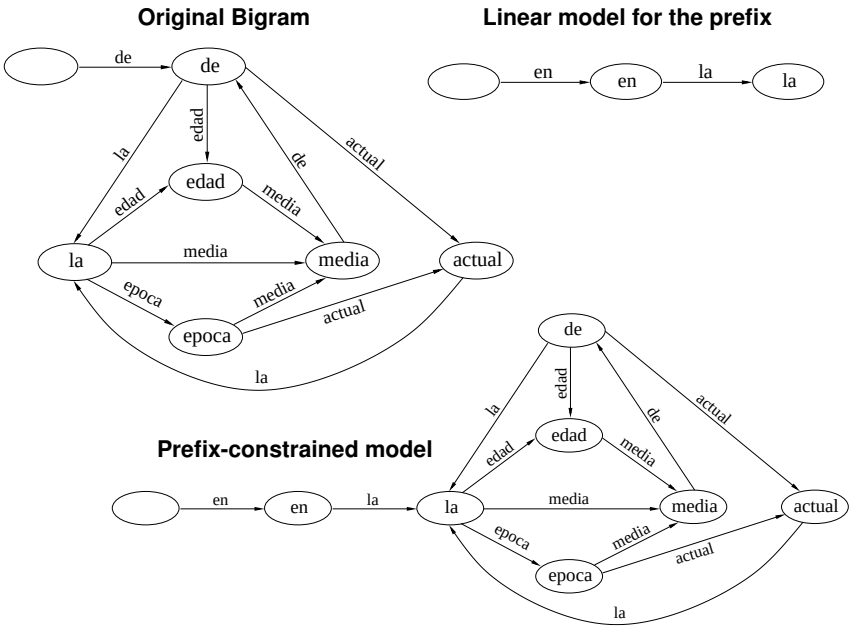
$$P(s \mid p) \approx \prod_{i=k+1}^{k+N-1} P(w_i \mid w_{i-N+1}^{i-1}) \cdot \prod_{i=k+N}^{l} P(w_i \mid w_{i-N+1}^{i-1})$$

CATTI “Dynamic” Language Modeling

Training samples

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

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 define the prefix “*en la*”.

CATTI Search

- Easily solved by the Viterbi algorithm
- A Viterbi computation is neaded in each CATTI interaction: Computational cost grows quadratically with the number of interactive steps.
- Computing cost can be somewhat alleviated by a CATTI-specific implemenmtation of Viterbi decoding
- Alternative solution (details later on): Compute a whole sentence decoding *word-graph* in the first step and then approach the following steps by searching precomputed solutions stored in the word-graphs.

HTR Corpora and Baseline Results

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

- 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-book
Writters		many	many	1
HMMs	Characters	78	80	78
	Tr. Ratio	2 779	808	460
Lang. Model	Lexicon	8 017	2 790	2 536
	OOV	921	518	1 400
	Tr. Ratio	128	4.4	2.5
WER (%)		**25.8**	**25.0**	**38.8**

Basic CATTI Results [Toselli et al., 2010]

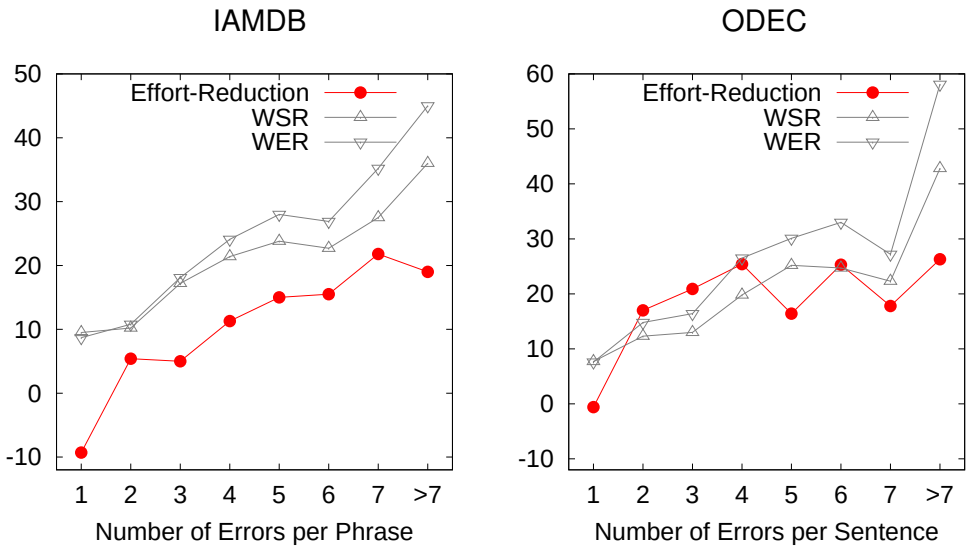
In all the experiments, only interaction at the *word level* is assumed; that is, each interaction step involves the correction of a *single, whole word* from the system-predicted suffix.

This precludes using the more practically-oriented KSR, but allows proper comparisons of the *estimated user effort* needed for non-interactive post-editing (WER) versus interactive processing (WSR).

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

Corpus	IAMDB	ODEC	CS (book)
WER (%)	25.8	25.0	38.8
WSR (%)	21.8	19.4	36.8
Effort-Reduction (%)	16	22	5

CATTI Results: impact of number of errors per sentence



CATTI performance for sentences with different number of errors

Word-graph (WG) CATTI decoding

Two implementations of the CATTI decoder:

- Repeated Viterbi decoding (baseline):
  - At each step a linear LM that accounts for the fixed prefix is "concatenated" with a standard LM to obtain a LM for $P(s \mid p)$
  - Computing cost grows quadratically with the number of words
- Faster approach based on WGs derived from the initial Viterbi decoding: Use this word-graph for interactive search to complete the prefixes accepted by the human transcriber
  - Efficient, linear computational cost
  - Drawback: some accuracy can be lost

Word-graph based plain (not interactive) Viterbi-like search

Let w be a word sequence and let $\mathcal{P}(w)$ be the set of all the paths in the WG associated with w; i.e., all the edge sequences $e_1, e_2, \ldots, e_l$ such that $w = \omega(e_1), \omega(e_2), \ldots, \omega(e_l)$. The most probable word sequence is computed as:

$$\hat{w} = \arg\max_{w} P(w) \qquad P(w) = \sum_{e_1, e_2, \ldots, e_l \in \mathcal{P}(w)} \prod_{k=1}^{l} p(e_k)$$

where the probability of an edge $e = (i, j)$ is:

$$p(e) = P(x_{t(i)}^{t(j)} | \omega(e)) \cdot P(\omega(e))$$

By approximating the sum to compute $P(w)$ with the maximum, $\hat{w}$ can be easily computed by Dynamic Programming:

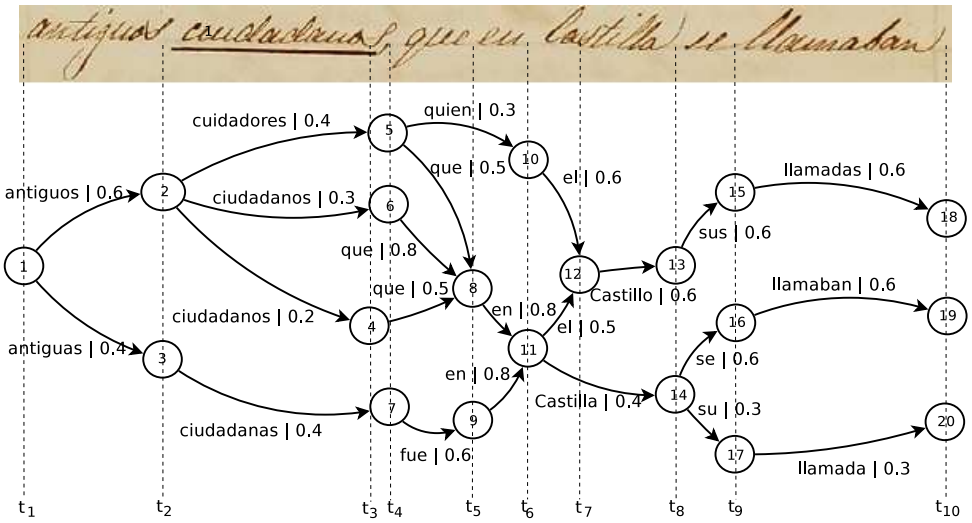
$$\hat{w} \approx \arg\max_{w} \max_{e_1, e_2, \ldots, e_l \in \mathcal{P}(w)} \prod_{k=1}^{l} p(e_k) = \arg\max_{e_1, e_2, \ldots, e_l \in \mathcal{P}} \prod_{k=1}^{l} p(e_k)$$

where $\mathcal{P}$ is the set of all the paths of the WG.

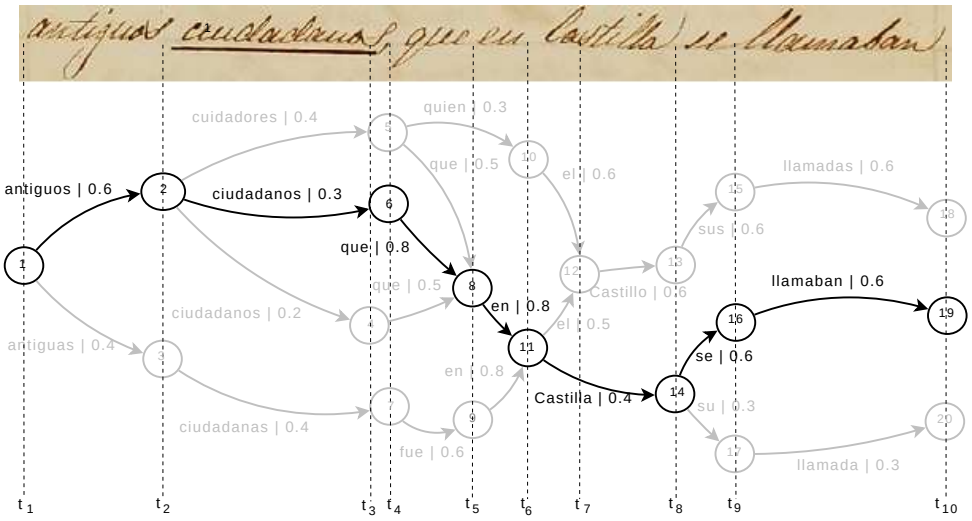
Instead of $P(e)$, GSF and WIP balanced log-prob scores, $\varphi(e)$, are used in practice:

$$\varphi(e) = \log P(x_{t(i)}^{t(j)} \mid \omega(e)) + \alpha \log P(\omega(e)) + \beta$$

Word-graph example



Best path in the example WG and the corresponding $\hat{w}$



$\hat{w}$ = "antiguos ciudadanos que en Castilla se llamaban"

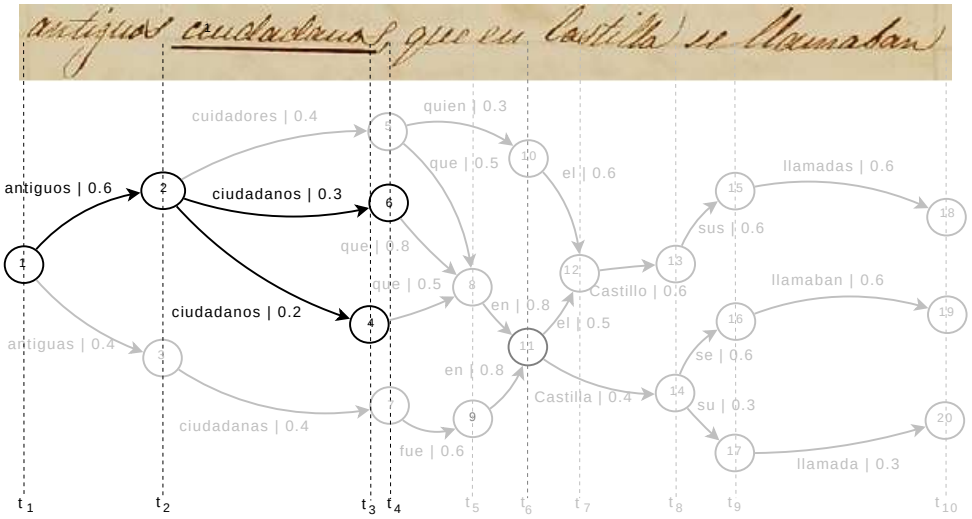
Word-graph based interactive-predictive search

- If the initial best word sequence, $\hat{w}$, is not correct, the user amends it from left to right, defining a correct prefix p
- The decoder parses p on the WG defining a set of nodes Q_p corresponding to paths from the initial node whose associated word sequence is p
- Departing from the nodes in Q_p, the decoder continues searching for a suffix s that maximizes the posterior probability:

$$\hat{s} = \arg\max_{s} \max_{q \in Q_p} P(x_1^{t(q)} \mid p) \cdot P(x_{t(q)+1}^{M} \mid s) \cdot P(s \mid p)$$

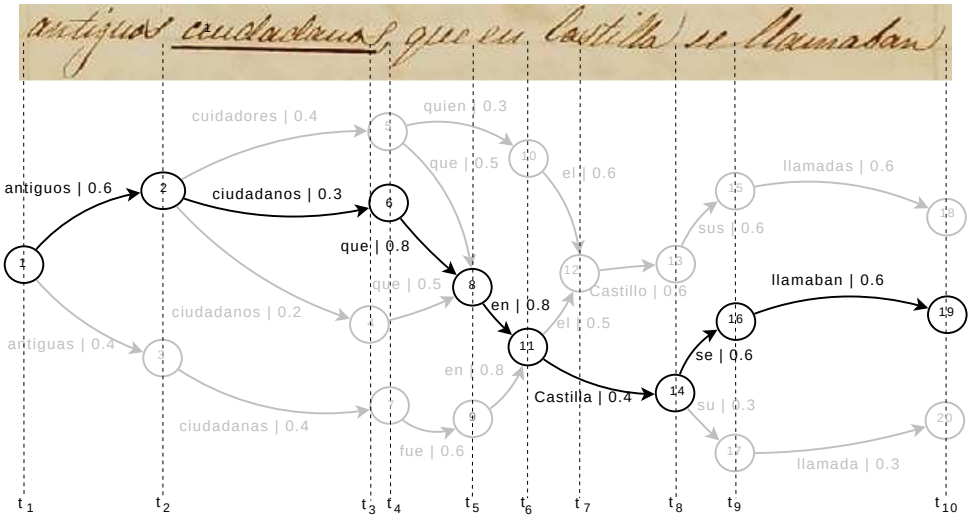
$P(x_1^{t(q)} \mid p)$, $P(x_{t(q)+1}^{M} \mid s)$ (HMM probabilities) and $P(s \mid p)$ (prefix-conditioned N-Gram probabilities) are trivially computed from the scores of the WG edges of p and s.

Word-graph based approach: Example



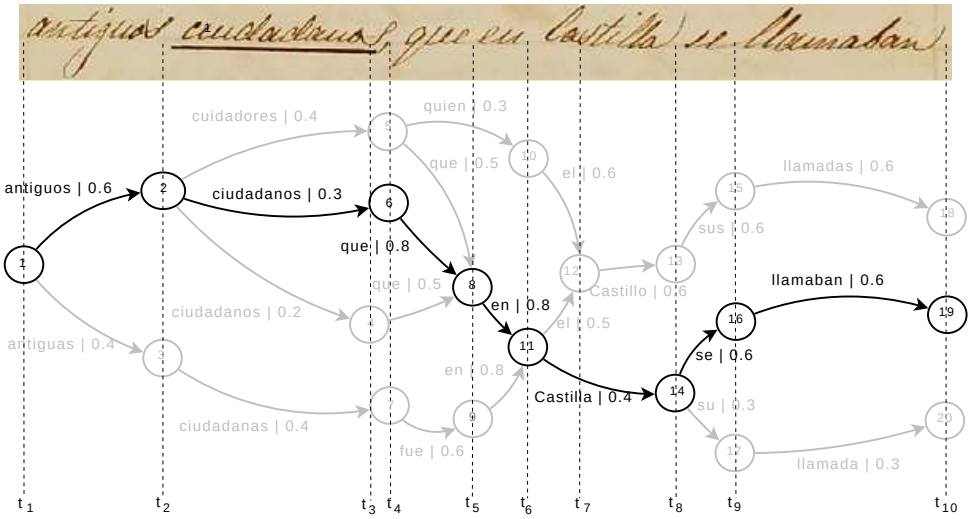
$p =$ “antiguos ciudadanos”

Word-graph based approach: Example



$p =$ “antiguos ciudadanos”

Word-graph based approach: Example



$p =$ "antiguos ciudadanos"

Problem: what happnes if p is not exactly an initial path in the WG?

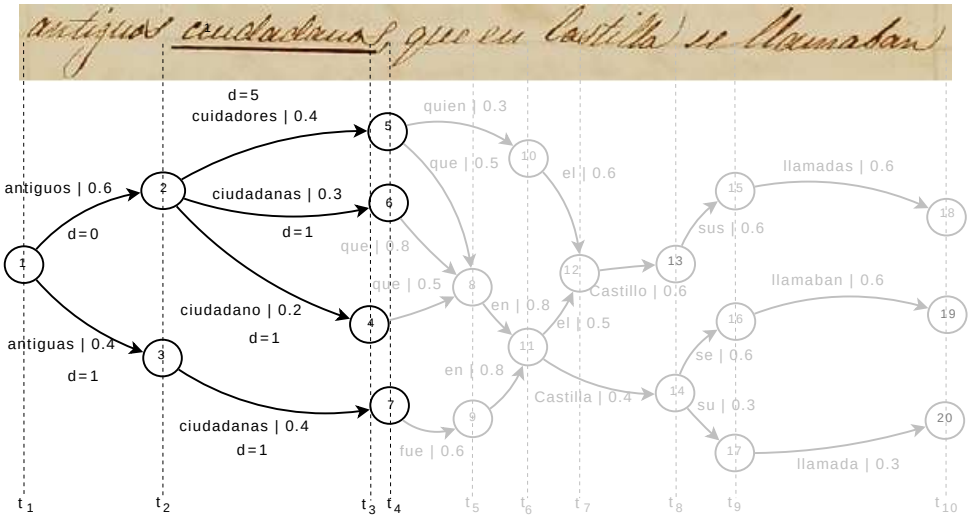
Word-graph Error-Correcting Prefix Parsing

- Solution: do not use the given prefix p but look for a prefix $\hat{p}_e$, among all the possibles initial paths in the word graph, which best matches p
- This is essentially the same Error Correcting Parsing process explained before, but now the optimization must find also the (unknown) best suffix. After some manipulations, we get:

$$(\hat{p}_e, \hat{s}) \approx \arg\max_{p_e, s} \max_{q \in Q} P(x_1^{t(q)} \mid p_e) \cdot P(x_{t(q)+1}^{M} \mid s) \cdot P(s \mid p_e) \cdot P(p_e \mid p)$$

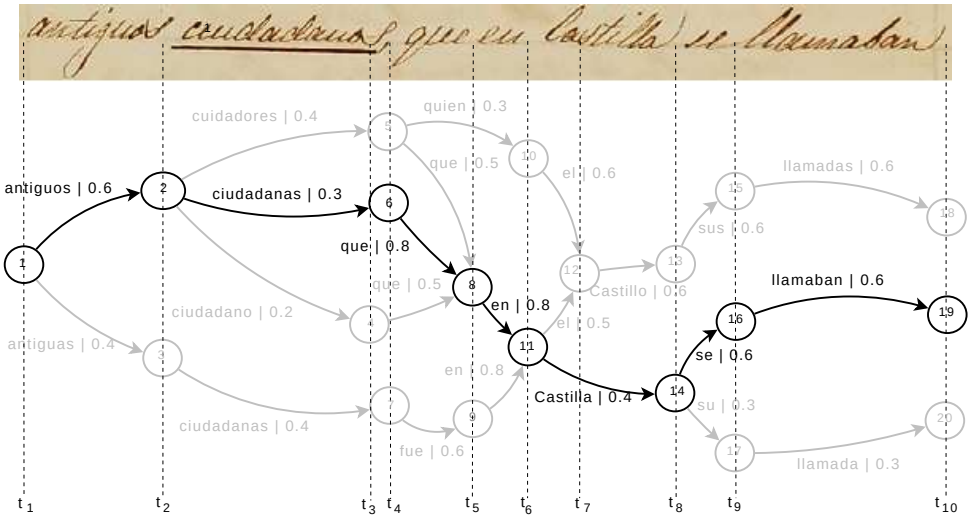
where, as before, $P(p_e \mid p)$ measures the similarity between p_e and p and is modelled as an stochastic edit distance based on the character-level lexical differences between the words.

Word-graph Error-Correcting Parsing: Example



$p =$ “antiguos ciudadanos”

Word-graph Error-Correcting Parsing: Example



$p =$ “antiguos ciudadanos”

CATTI results using Word-Graphs

Performance of non-interactive off-line HTR and CATTI along with the relative difference between them (EFR).

Corpus	WER	WSR	EFR
IAMDB	25.3	22.5	11.1
ODEC	22.9	21.5	6.1
CS-book	33.5	32.3	3.6

- EFR is still positive but, as compared with the direct Viterbi approach, EFR values are significantly worse.
- But computing cost is now very much lower and allows for other, more ergonomic interaction modes based on weaker feedback. In the end, this lead to a net overall EFR improvement.

[1] All the results in this section are slightly better than those appearing previously, thanks to preprocessing and modelling improvements.

Weaker feedback: Pointer Actions (“Clicks”)

- CATTI interaction: a pointer-action (PA) to position the cursor + typing the correct word
- The PA explicitly indicates a suffix the user is not happy with. It can trigger the system to propose an alternative suffix
- The system changes the suffix with the next most probable suffix whose first word is different
- Additiona PA (“clicks”) can be issued if the first one fails to provide the expected suffix
- The first PA, called “Single PA” (S-PA), does not involve extra human effort
- Many explicit user corrections are avoided
- A very low interaction system response times is now crucial.

Weaker feedback: Example

	x							
STEP-0	**p**							
STEP-1	**ŝ**	**antiguos**	**cuidadores**	**que**	**en**	**el Castillo**	**sus**	**llamadas**
	m		↑					
	p’	**antiguos**						
	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -
	ŝ		**cortesanos**	**que**	**en**	**el Castillo**	**sus**	**llamadas**
	v		ciudadanos					
	p	**antiguos**	**ciudadanos**					
STEP-2	**ŝ**			**que**	**en**	**el Castillo**	**sus**	**llamadas**
	m					↑		
	p’	**antiguos**	**ciudadanos**	**que**	**en**			
FINAL	**ŝ**					**Castilla**	**se**	**llamaban**
	v							**#**
	p $\equiv$ **T**	**antiguos**	***ciudadanos***	**que**	**en**	**Castilla**	**se**	**llamaban**

First hypotesis word errors: 5 $WER = 71\%$ Basic CATTI interactions: 2 $WSR = 29\%$

Number of user interactions using S-PA CATTI: 1 $WSR = \frac{1}{7} \cdot 100 = 14\%$

Weaker Feedback: Formulation and Modelling

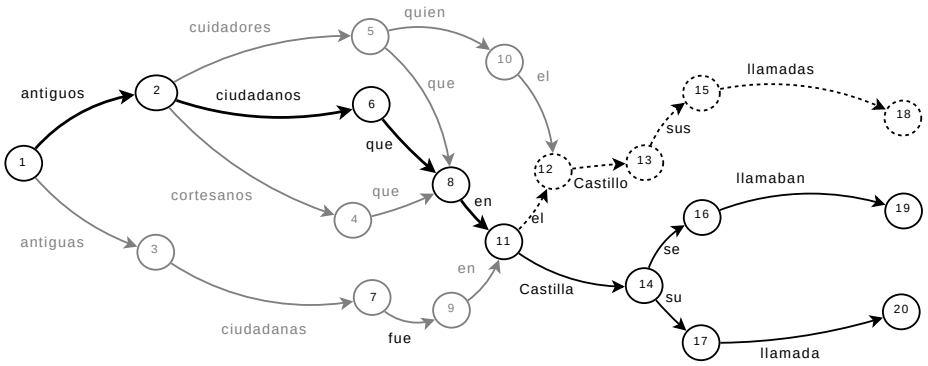
- Te decoder has information from the input image x, the validated prefix p' and the erroneous word e that follows the validated prefix:

$$\hat{s} \;=\; \arg\max_s P(s \mid x, p', e) \;\approx\; \arg\max_s P(x \mid p', s) \cdot P(s \mid p', e)$$

  - $P(x \mid p', s)$: is modeled using HMMs, following similar assumptions and developments carried out previously.
  - $P(s \mid p', e)$: can be provided by a language model constrained by the validated prefix p' and the erroneous word e that follows it.

Weaker feedback: Searching

- Using the Viterbi algorithm: A special language model can be built modifying the “Preffix-conditioned Language Model“ so that the errorneous word has null probability
- Simpler and much faster implementation using word-graphs: assign null probability to the edge labelled with the word e after matching the prefix



Weaker feedback results: Comparing estimated effort reduction

	Viterbi	Word-graph	
Corpus	EFR	EFR	EFR-S
IAMDB	16.6	11.1	26.5
ODEC	17.5	6.1	20.5
CS-book	4.2	3.6	15.2

- All EFR figures are percentages
- The WG approach for for CATTI + single PAs leads to better EFR (EFR-S) than with the direct Viterbi aproach (without PAs)
- These results are obtained with very low computing cost
- Further EFR by means of (up to 3) additional clicks

Conclusions

- Current HTR accuracy is not enough for fully automatic high quality transcription of most handwritten text images of interest
- Human post-editing can be very expensive and hardly acceptable by profesional transcribers (paleographers, e.g.)
- *Computer Assisted, Interactive-Predictive processing* offers promise for *significant improvements in practical performance and user acceptance*

Bibliography

- F. Jelinek. “Statistical Methods for Speech Recognition”. MIT Press, 1998.
- I. Bazzi, R. Schwartz, J. Makhoul. “An Omnifont Open-Vocabulary OCR System for English and Arabic”. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) Vol.21 pp.495-504, 1999.
- A. Vinciarelli, S. Bengio, H. Bunke. “Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models”. IEEE Trans. on PAMI, Vol.26, pp.709-720, 2004.
- A. H. Toselli, A. Juan, D. Keysers, J. Gonzlez, I. Salvador, H. Ney, E. Vidal and F. Casacuberta. “Integrated Handwriting Recognition and Interpretation using Finite-State Models”. Int. Journal of Pattern Recognition and Artificial Intelligence, 18(4):519-539, June 2004.
- E. Vidal, L. Rodriguez, 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.
- 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.
- A.H. Toselli, V. Romero, M. Pastor and E. Vidal. “Multimodal interactive transcription of text images”. Pattern Recognition, Vol.43, N.5, pp.1814–1825, April 2010.
- V.Romero, A.H.Toselli and Vidal: “Multimodal Interactive Handwritten Text Transcription”, Vol. 80. World Scientific, 2012.

Demonstrations of Interactive-Predictive (CATTI) HTR Prototypes

From TRANSCRIPTORIUM:

http://transcriptorium.eu
→ DEMONSTRATIONS → Computer Assisted Text Transcription (UPVLC)

- PLANTAS: XVII c. Spanish manuscript on Botany
- HATTEM: XV c. Dutch manuscript on Medecine
- Bentham collection XVIII-XIX c. English manuscripts on Philosophy & Law

Others:

https://www.prhlt.upv.es/showcase/htr
→ Demonstrations → IHT prototype