Active Learning in Handwritten Text Recognition using the Derivational Entropy

Verónica Romero, Joan Andreu Sánchez, Alejandro H. Toselli
Pattern Recognition and Human Language Technology Research Center,
Universitat Politècnica de València, Camino Vera s/n, 46022, Valencia, Spain

Introduction

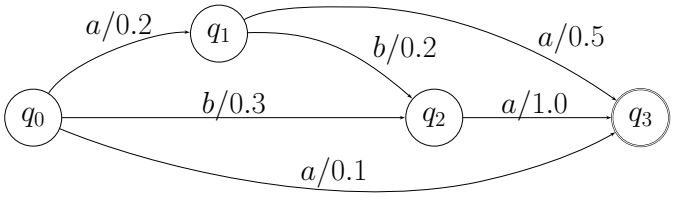
- State-of-the-art HTR systems are based on statistical models.
- These models need large corpora for training, consisting in text line images with their corresponding transcripts.
- The manual annotation is expensive.
- Alternative: Active Learning techniques to selecting the most informative samples.
- The derivational entropy computed from word graphs cab be used for assessing how informative an unannotated trainin sequence is.

Probabilistic Finite-State Automata and Word Graphs normalization

- A word graph (WG) is a labelled weighted directed acyclic graph:

$$\mathcal{G} = \langle Q, \Sigma, \delta, I, F, P \rangle$$

- During the decoding proces a WG is obtained for each text line image.
- A WG reperesents the transcriptions with higher probability of the given image.
- The obtained WG after the decoding process is not guaranteed to be proper nor consistent.
- In order to compute derivational entropy a WG has to be normalized.



- To transform a WFA into a proper PFA, all transitions weights $W(q, v, q') \in \delta$ must be multiplied by $\mathcal{N}(q')/\mathcal{N}(q)$ with $\mathcal{N}(q_f) = 1$.
- The normalition vector $\mathcal{N}$ is defined as:

$$\mathcal{N} = (I - M)^{-1}\nu$$

where

  - final vector ν is $\nu(q) = \sum_{v\in\Sigma} W(q, v, q_f)\ \forall q \in Q - \{q_f\}$
  - the characteristic matrix M is $M(q, q') = \sum_{v\in\Sigma} W(q, v, q')$.

- $\mathcal{N}$ can be efficiently computed in an acyclic graph as:

$$\mathcal{N}(q) = \sum_{i=0}^{|Q|-2} \mathcal{W}_{\mathcal{A}}(\Theta_{\mathcal{A}}(q, q_f, i+1))$$

where $\Theta_{\mathcal{A}}(q', q, l)$ is defined as the set of all paths starting in q', and ending in q of length l.

Derivational Entropy of an Acyclic PFA

Definition 1. *The derivational entropy of a PFA $\mathcal{A}$ is defined as:*

$$H(\mathcal{A}) = \sum_{\theta\in\Theta_{\mathcal{A}}(q_0, q_f, \cdot)} -p_{\mathcal{A}}(\theta) \log p_{\mathcal{A}}(\theta)$$

Considering an acyclic PFA with the states nominated in topological order, the derivational entropy can be efficiently computed as:

$$H(\mathcal{A}) = -\sum_{(q,v,q')\in\delta} P(q, v, q') \log P(q, v, q') \sum_{i=0}^{|Q|-2} \alpha_{\mathcal{A}}(i, q)$$

ESPOSALLES corpora

- Marriage license book from the 17th century
- Written by one writer in old Catalan
- Composed by 173 pages that contain 5,437 lines

Experimental results

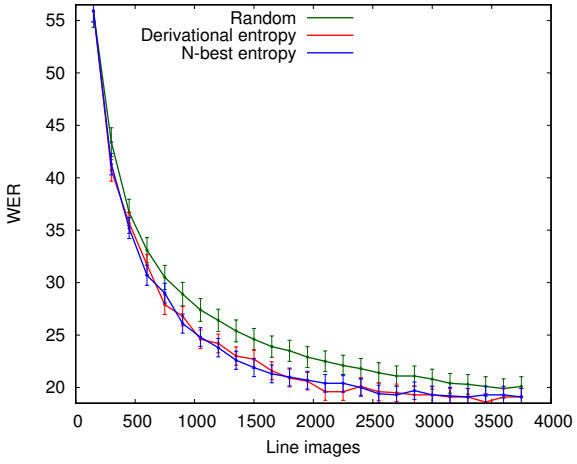
- An initial model ϕ is trained using a small set of annotated samples $\mathcal{L}$.
- The informativeness of each sample in a set of unannotated samples $\mathcal{U}$ is measured by the derivational entropy.
- A set $X_B^* \subseteq \mathcal{U}$ of B unannotated samples are selected according to the result of the entropy and annotated by an expert.
- The set $\mathcal{L}$ is updated with the new annotated samples and they are removed from $\mathcal{U}$.
- This iterative process continues until some stopping criterion.

System Setup

- The HTR system was based in HMMs and n-grams.
- The *Word Error Rate* is used to assess the quality of the transcription.

Results

	$\mathcal{L}$	$\mathcal{U}$	Test	Total
Lines	150	4,470	827	5,447
Running words	1,650	50,234	8,893	60,777
Vocabulary	397	2,914	1,119	3,500



- The batch size B was 150 lines
- The random experiment was repeated 10 times

Conclusions

- The efficient coputation of the derivational entropy in an acyclic WG has been presented.
- The derivational entropy can be used as confidence measure for AL with good results.
- An efficient WG normalization has been also introduced.

Acknowledgements

This work has been partially supported through the European Union's H2020 grant READ (Recognition and Enrichment of Archival Documents) (Ref: 674943) and the MINECO/FEDER, UE project TIN2015-70924-C2-1-R