tranScriptorium

An intelligent sample selection approach to language model adaptation for hand-written text recognition

Jafar Tanha, Jesse de Does, and Katrien Depuydt

Instituut voor Nederlandse Lexicologie (INL),

The Netherlands

INL

Outline

- Linguistic resources for HTR
- General Setting of the Resources
- Current Approaches
- Main Challenge
- Proposed method and Algorithm
- Datasets and Experiments
- Conclusion

Linguistic resources for HTR

Main issue of language modeling for HTR in case of **Historical** *data:*

- *Data sparsity due to Nonstandardized language*
  - (Historical) spelling variation (→ unknown word problem)
  - Limited amount of relevant corpus material

Countermeasures:

- Develop normalization strategies using variation lexica

- **Try to use a combination of in-domain and general corpora, domain adaptation.**

Combination of in-domain and out-of-domain resources

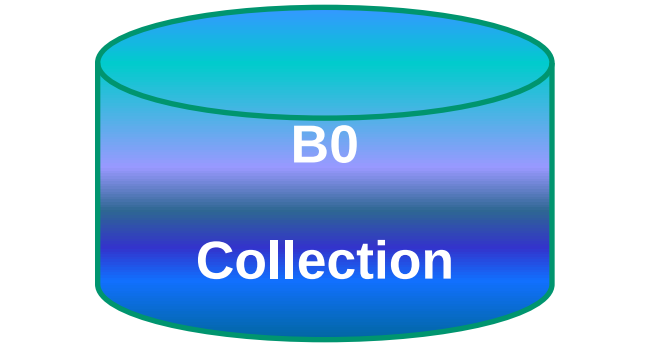
Focus for this presentation:

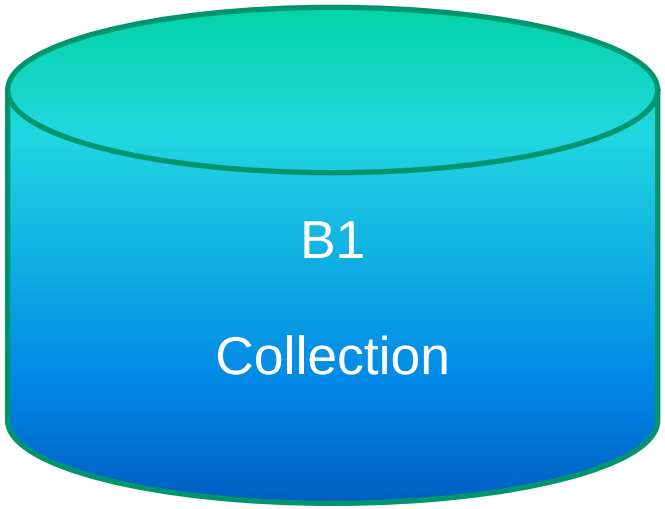
- How to combine and use in-domain and out-of-domain resources to improve the performance of the HTR system

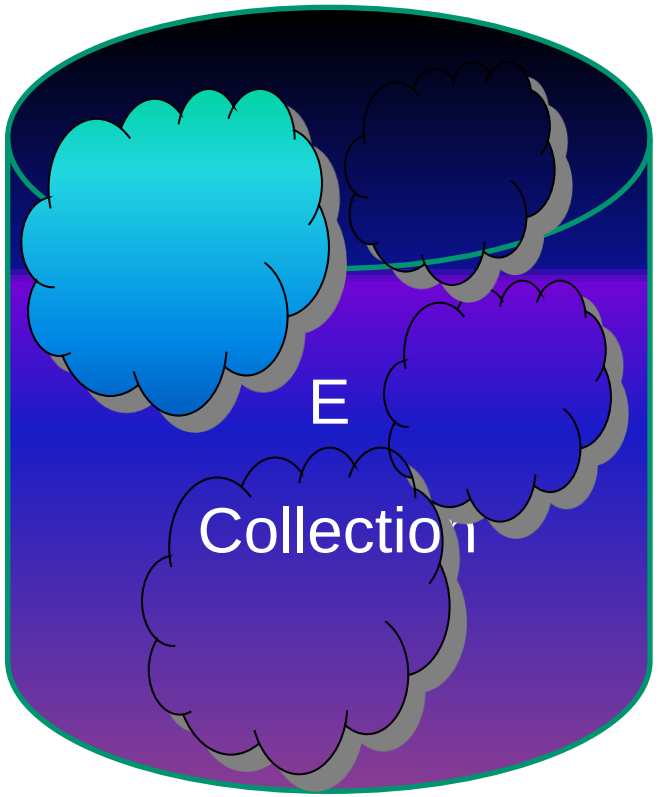
- One can concatenate the resources, however:
  - indiscriminate use of *out-of-domain* data may not benefit, in fact even deteriorate system
  - the use of the complete out-domain data for training may increase the complexity of the system, making the decoding process almost intractable

- We consider the problem as a *domain adaption* problem for the HTR system, to be approached by ***intelligent sample selection***

General resource situation for Sample Selection







$|B_0| < |B_1| < |E|$, and the B_1 material is (much) more similar to the B_0 material than the material from E. B_0 and B_1 may arise from a partition of a corpus B.

Basic procedure

The basic procedure to combine resources is:

- Obtain *relevant* subset from the out-of-domain data
- Build for each resource a language model (LM).
- Next, *interpolate* the resulting LMs
  - For example, for two models LM_1 and LM_2, The interpolated model LM_λ is defined by

$$p_\lambda(w|h) = \lambda\, p_1(w|h) + (1-\lambda) p_2(w|h)$$

Where the interpolation parameter λ in [0, 1].

The main challenges

- Ranking *criteria* for relevance of out-of-domain data
- Selection *procedure*
  - What: Sentences, documents, lines, …
  - How: ranking, stochastic sampling, …?
- *Combination scenario*
  - Use subset of out-of-domain and interpolate them
  - Use different subsets of out-of-domain and interpolate multiple models
  - …..

We propose an iterative method to select a set of informative resources instead of just using the complete out-of-domain data.

Ranking criteria

- Use the entropy difference [Moore e.a. 2010]:

$$H_{in\text{-}domain}(s) - H_{out\text{-}of\text{-}domain}(s)$$

- Issue: the perplexity/entropy cannot be a proper criterion when the number of OOVs for each model differs
  - In general: A model with a larger vocabulary which in practice may perform better when deployed in the HTR system, can end up having worse perplexity than one with a smaller vocabulary, estimated from a smaller training corpus

Ranking criteria: problems

Sentence from Hattem: OP DIE KINNEBACKE OFMEN SALT PLAESTEN

LM estimated from ~40 pages	**LM estimated from ~300 pages**
p(OP | <s>) = [2gram] 0.00158299 [-2.80052]	p(OP | <s>) = [2gram] 0.00592399 [-2.22739]
p(DIE | OP ...) = [2gram] 0.0759172 [-1.11966]	p(DIE | OP ...) = [2gram] 0.109495 [-0.960605]
p(<unk> | DIE ...) = [OOV] 0 [-inf]	p(<unk> | DIE ...) = [OOV] 0 [-inf]
p(<unk> | <unk> ...) = [OOV] 0 [-inf]	p(OFME | <unk> ...) = [1gram] 3.51984e-05 [-4.45348]
p(SALT | <unk> ...) = [1gram] 0.000709825 [-3.14885]	p(SALT | OFME ...) = [1gram] 0.000305239 [-3.51536]
p(PLAESTEN | SALT ...) = [1gram] 0.00032598 [-3.48681]	p(PLAESTEN | SALT ...) = [1gram] 8.90856e-05 [-4.05019]
p(</s> | PLAESTEN ...) = [2gram] 0.0405917 [-1.39156]	p(</s> | PLAESTEN ...) = [2gram] 0.278732 [-0.554813]
1 sentences, 6 words, 2 OOVs	1 sentences, 6 words, 1 OOVs
0 zeroprobs, logprob= -11.9474 ppl= 245.177 ppl1= 970.176	0 zeroprobs, logprob= -15.7618 ppl= 423.616 ppl1= 1420.26

Proposed methods for ranking

We take into account the PPL and OOVs in our formulation:

- Additive criterion
  (log *PPL)* + |*OOV*|/|V|
- Multiplicative criterion
  (log PPL) * |*OOV*|/|*V*|
- Average word probability, with *p(unknown)* set to 0
  (1/log PPL)(1-* |*OOV*|/|*V*|*)*

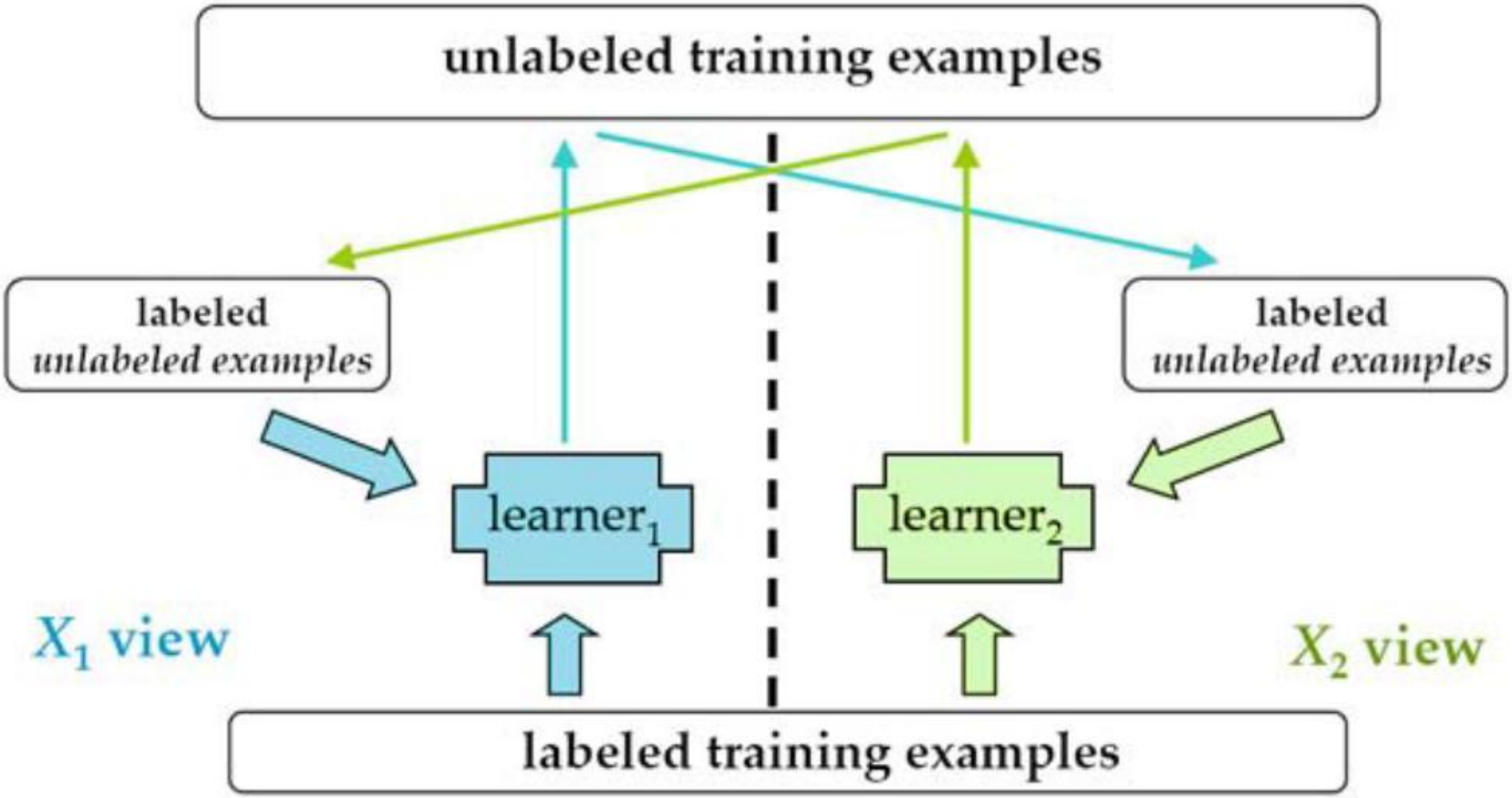
Approaches to sample selection

- In [Gao et. al. 2000], a text retrieval approach has been proposed for domain adaptation problem. The main idea of this approach is to avoid items specific to the out-of-domain data by removing n-grams likely to be infrequent in new documents, based on a partitioning of the training data.
- Moore and Lewis [Moore & Lewis 2010] have proposed a cross-entropy based approach to sample randomly from the out-of-domain data using perplexity as a main criterion.
- Gascó et.al. [Gascó et. al. 2012] address a data selection approach from out-of-domain corpora by approximating the probability of an in-domain corpus.

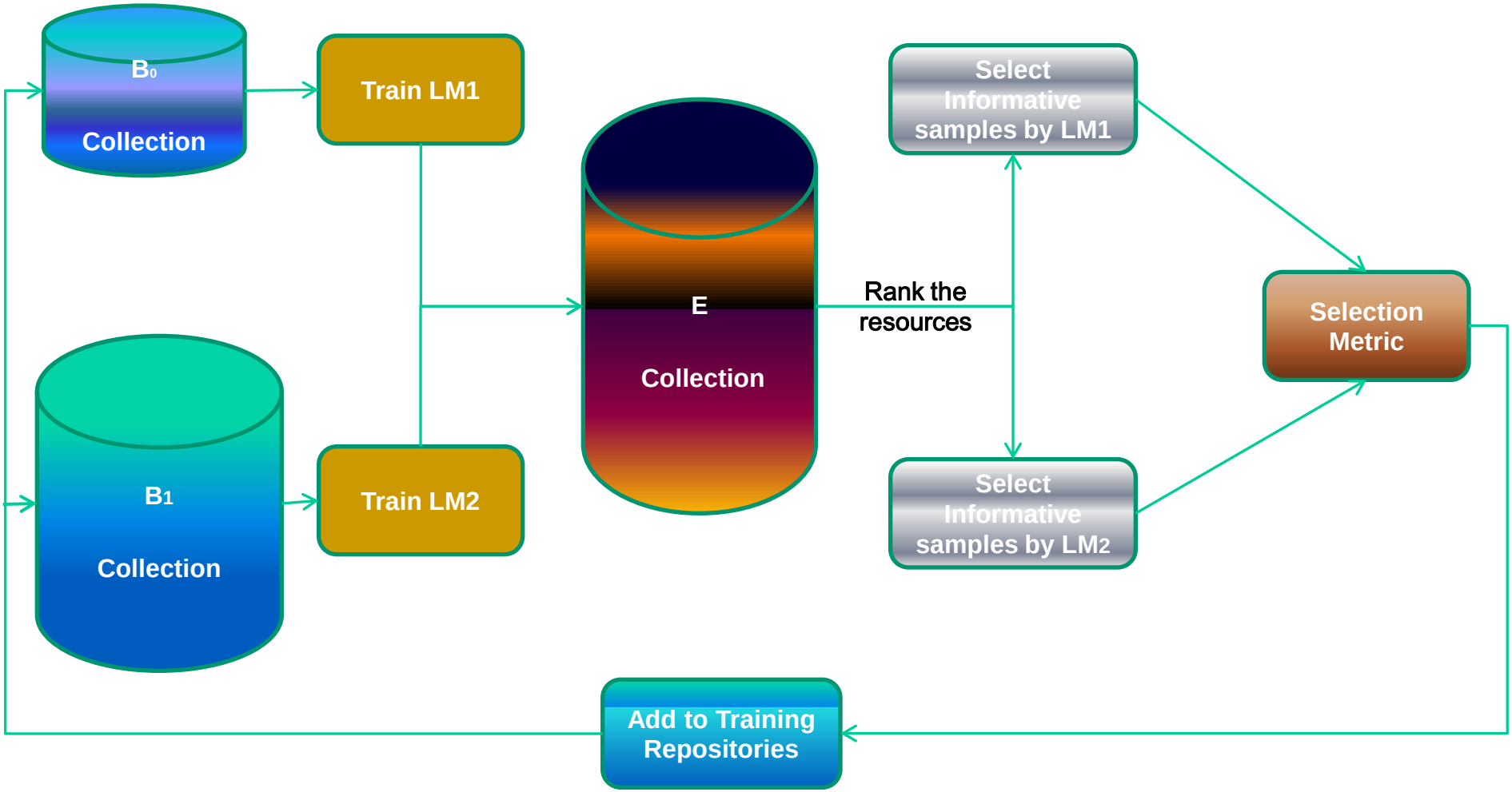
Our approach (1)

- Inspired by *Co-training*
- Co-training gradually exploits informative unlabeled data and assigns labels to them
- in Co-training two or more classifiers are trained from iteratively growing training sets
- We fit our problem in this framework and propose two iterative approaches.
- We consider each language model, which has been trained on an in-domain subcorpus, as a classifier and use it for ranking the out-domain data.

Co-training Approach



Our approach (1)



The Agree-Co Algorithm

Algorithm for domain adaptation *Agree-Co*(B_0, B_1, E, *max_iterations*, *threshold*)

Initialize:

t=0 , *conf*=0;

select a resource scoring criterion C

Begin

While (t < *max_iterations* && *conf* < *threshold*)

Begin

Build LM_0 and LM_1 from B_0 and B_1

For each R_i in E **do**

Begin

Compute $C_0(R_i)$ and $C_1(R_i)$

End

H_0 := a high-confidence subset of E, selected by best values w.r.t. C_0

H_1 := a high-confidence subset of E, selected by best values w.r.t. C_1

Assign a value to *conf* by averaging the worst values of C_0 and C_1 on H_0 resp. H_1

$S_t := H_0 \cap H_1$

$B_0 := B_0 \cup S_t$

$B_1 := B_1 \cup S_t$

$E := E \setminus S_t$

t := t+1

End // while

Output: Selected Resources for domain adaptation and B_0 and B_1.

End/ / Algorithm

Algorithm 4-1, pseudo-code of the Agree-Co Algorithm

Corpus Data (Bentham & ECCO)

- The Bentham corpus of transcribed manuscripts (about 15.000 pages and 5M words).
- The public part of the ECCO (Eighteenth Century Collections Online) corpus, about 70M words.

With these two corpora, we make a two-level *in-domain/out-of-domain* distinction:

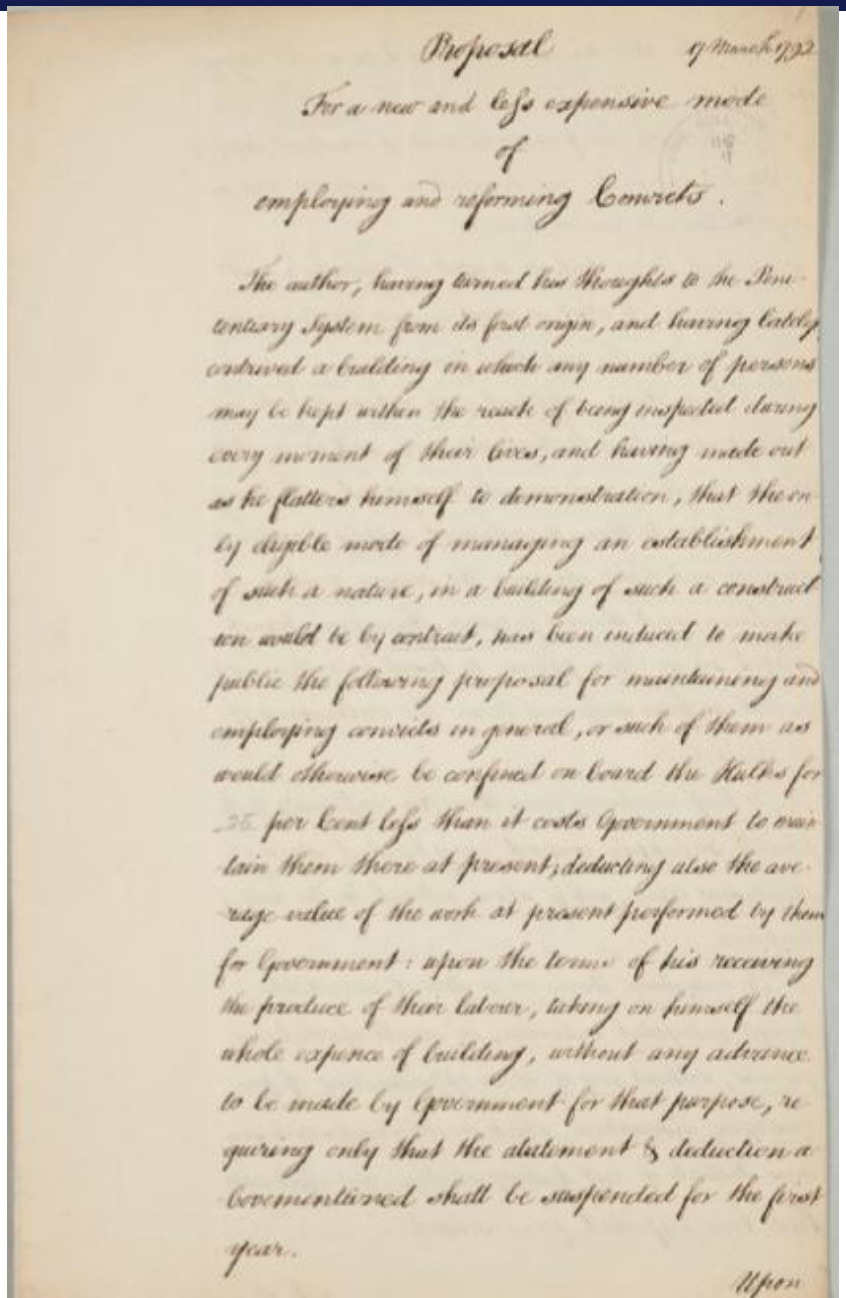
- The ECCO corpus is considered as an out-of-domain resource.

- Within the set of Bentham transcripts we distinguish: the set of “Batch 1” ground truth transcriptions (~400 pages) as an in-domain resource and the rest as out-of-domain.

Datasets

Resources	Size	Function
Bentham In “Batch 1”	57.7 (kb)	LM training, HTR training
Bentham Out	38.3(mb)	LM training
ECCO	425.8(mb)	LM training
Test set	6.5 (kb)	HTR testing

Data (Bentham)



Proposal 9 March 1792

For a new and less expensive mode

of

employing and reforming Convicts.

The author, having turned his thoughts to the Penitentiary System from its first origin, and having lately contrived a building in which any number of persons may be kept within the reach of being inspected during every moment of their lives, and having made out as he flatters himself to demonstration, that the only eligible mode of managing an establishment of such a nature, in a building of such a construction would be by contract, has been induced to make public the following proposal for maintaining and employing convicts in general, or such of them as would otherwise be confined on board the Hulks for 25 per Cent less than it costs Government to maintain them there at present; deducting also the average value of the work at present performed by them for Government: upon the terms of his receiving the produce of their labour, taking on himself the whole expence of building, without any advance to be made by Government for that purpose, requiring only that the abatement & deduction above mentioned shall be suspended for the first year.

Upon

Experiments (1)

We have applied the following scenarios for interpolation:

- Combining two Bentham resources (In and Out) and using a dictionary from the merged data to train the LM (Merged-InOut-Dic-InOut).
- Interpolating Bentham In and Out domain resources using dictionary from In domain data (Inter-InOut-Dic-In).
- Interpolating Bentham In and Out domain resources using dictionary from both In and Out domain data (Inter-InOut-Dic-InOut).
- Combining two Bentham resources and interpolate the resulting LM with the LM of ECCO using dictionary from the merged data to train the LM (Inter-In+OutECCO-Dic-InOutECCO).
- Interpolating Bentham In and Out domain resources with ECCO collection using dictionary from Bentham In and Out domain data (Inter-InOutECCO-Dic-InOut).
- Interpolating Bentham In and Out domain resources with ECCO collection using dictionary from all of them (Inter-InOutECCO-Dic-InOutECCO).

Experiments(2)

We have applied the following scenarios for sample selection algorithm (AgreeCo):

- Single Iteration: we select the best 15% of the high confidence resources using the proposed algorithms.
- Multiple Iterations: we set the number of iterations to 20 iterations.

Results (1)

Method	WER %	WER without first word %	CER %	OOV %
Initial model using only Batch 1 training set	34.5	34.3	19.9	9.44
Merged-InOut-Dic-InOut	34.01	-	-	-
Inter-InOut-Dic-In	33.40	-	-	-
Inter-InOut-Dic-InOut	30.02	24.57	-	-
Inter-In+OutECCO-Dic-InOutECCO	31.7	26	16.5	-
Inter-InOutECCO-Dic-InOut	30.7	25.3	15.9	-
Inter-InOutECCO-Dic-InOutECCO	**28.31**	22.74	14.7	5.4

Results (2)

HTR results using Agree-Co*, with the additive criterion as selection metric.*

Method	WER %	WER without first word %	CER %
Agree-Co-Single	27.47	22.06	14.4
Agree-Co-Inter-Two	30.44	24.57	16.1
Agree-Co-Inter-Three	27.13	21.68	14.2

HTR system using Agree-Co*, with the multiplicative criterion as selection metric*

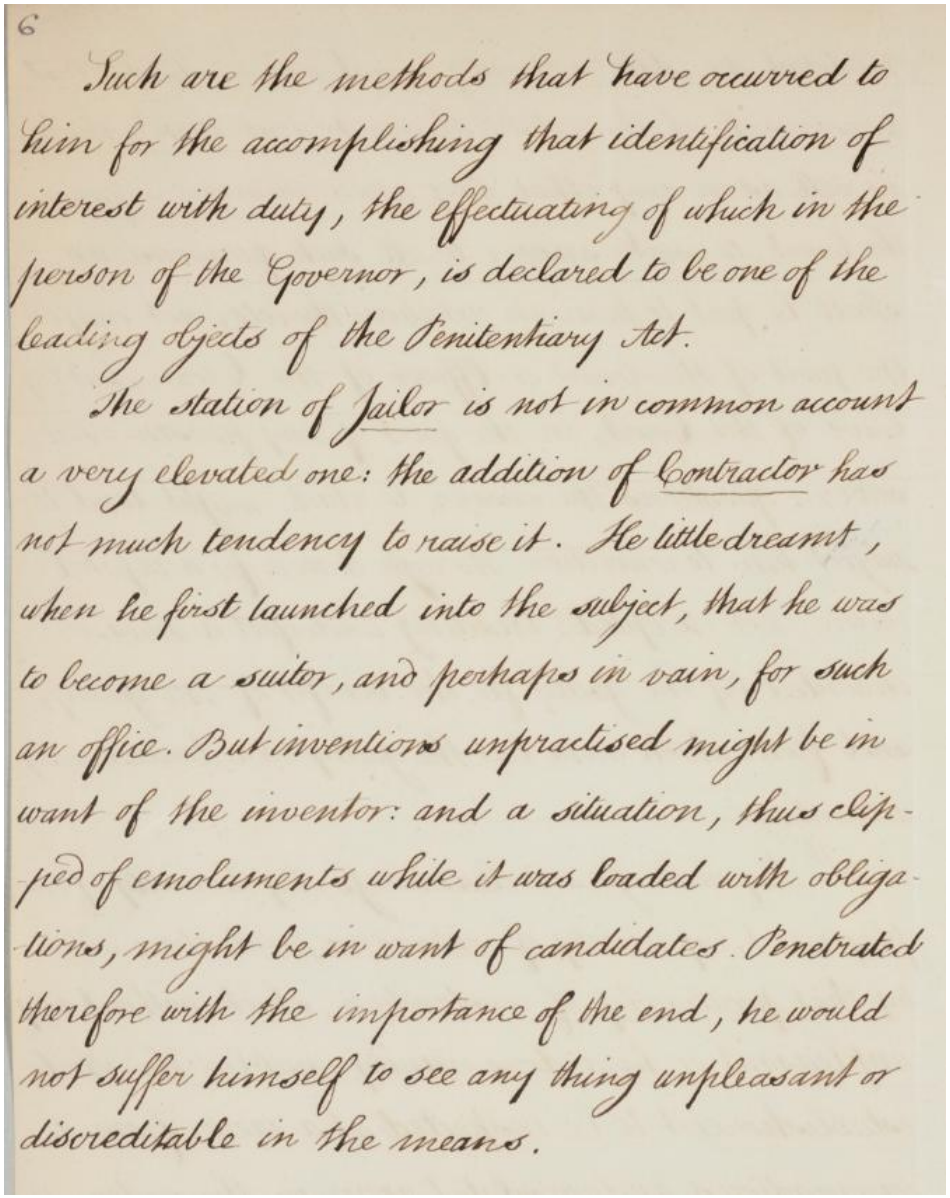
Method	WER %	WER without first word %	CER %
Agree-Co-Single	**27.04**	**21.68**	**14.1**
Agree-Co-Inter-Two	28.74	23.22	15.4
Agree-Co-Inter-Three	27.30	21.97	14.3

Results (3)

HTR system using Agree-Co*, with average word probability as selection metric*

Method	WER %	WER without first word %	CER %
Agree-Co-Single	28.02	23.12	14.9
Agree-Co-Inter-Two	29.51	23.99	-
Agree-Co-Inter-Three	**27.72**	**22.35**	**14.6**

Results (4)



6

Such are the methods that have occurred to him for the accomplishing that identification of interest with duty , the effectuating of which in the person of the Governor , is declared to be one of the leading objects of the Penitentiary Act .

The station of Jailor is not in common account a very elevated one : the addition of Contractor has not much tendency to raise it . He little dreamt , when he first launched into the subject , that he was to become a suitor , and perhaps in vain , for such an office . But inventions unpractised might be in want of the inventory : and a situation , this clip- -ped of emoluments while it was loaded with obliga- -tions , might be in want of candidates . Penetrated therefore with the importance of the end , he would not suffer himself to see any thing unpleasant or discreditable in the means .

Results (5)

6

F Such are the were tho also that have agreed to
him for the accompanied thing that indemnification of
interest with during the effectuating of which in the
person of the lower nor , is declared to be one of the
treading by acts of the Penitentiary Act
The situation of felon is not in common account
a very class acted one a the addition of Contract or has
not much under any to refuse it He little dream to
when The first touched into the subject that he was
to become a situation , and per happen arduum , for such
an office - But invented me improved used might be in
to cent of the i have whom and d situation , thus cloth
-panied of instaments whole it was be added with obliged
-tions might be in amount of and trades Samuel dated
therefore with the in pro hence of the end he would
not s infer himself to see any thing unpleasant or
did reducible in the threatens ,

6

ch are the methods that have occurred
in for the accomplishing that identification of
Arrest with duty , the effectuating of which in the
person of the Governor , is declared to be one of the
adding objects of the Penitentiary Act
he station of Jailor is not in common account
[a] very elevated one the addition of Contractor [has]
of much tendency to raise it He little dreamt
him The first launched into the subject , that he in
[to] become a suitor , and perhaps in vain , for such
[an] office But inventions unpractised might be [in]
art of the inventor outed All situation , thus clip
[-ped] of emoluments while it was loaded with obliged
little , might be in want of candidates . Penetrated
before with the importance of the end , he would
not suffer himself to see anything unpleasant [or]
creditable in the appears ,

Conclusions

- We have studied and tested several ways in which task-specific approaches to language modeling can improve handwritten text recognition results.
- Approaches to the combination of in-domain and out-of domain data have been shown to yield improvement in HTR performance.
- The proposed sample selection algorithm for domain adaptation outperforms the other general methods.

Future work

- Extend the work on sample selection for different datasets and combine it with elaboration of the approach to text normalization.
- Use the topic-modeling approach in order to sample selection.
- Use a clustering based approach in order to find similar data points to in-domain data and find an iterative way to combine more similar clusters.

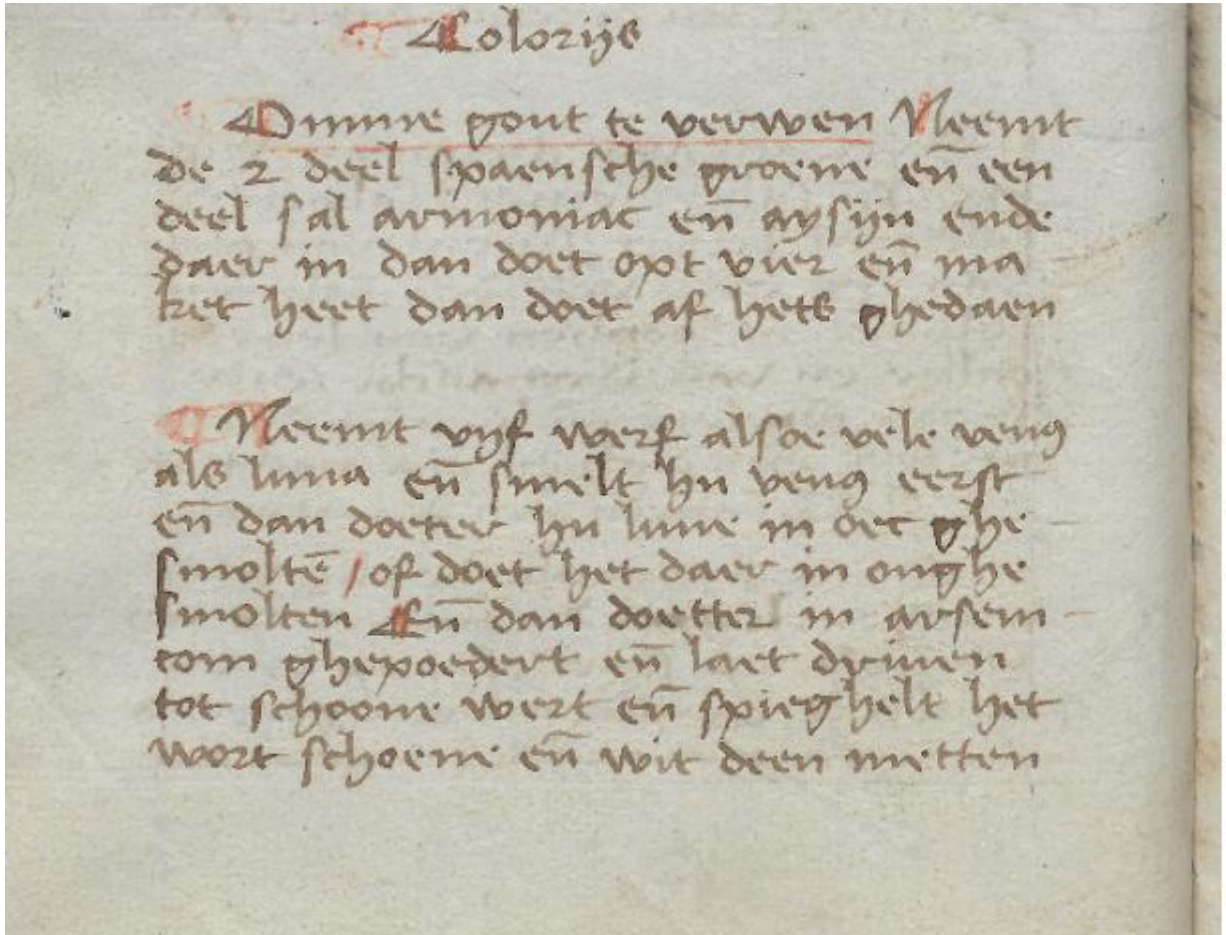
Thank you for your attention!

Any Question?

tranScriptorium

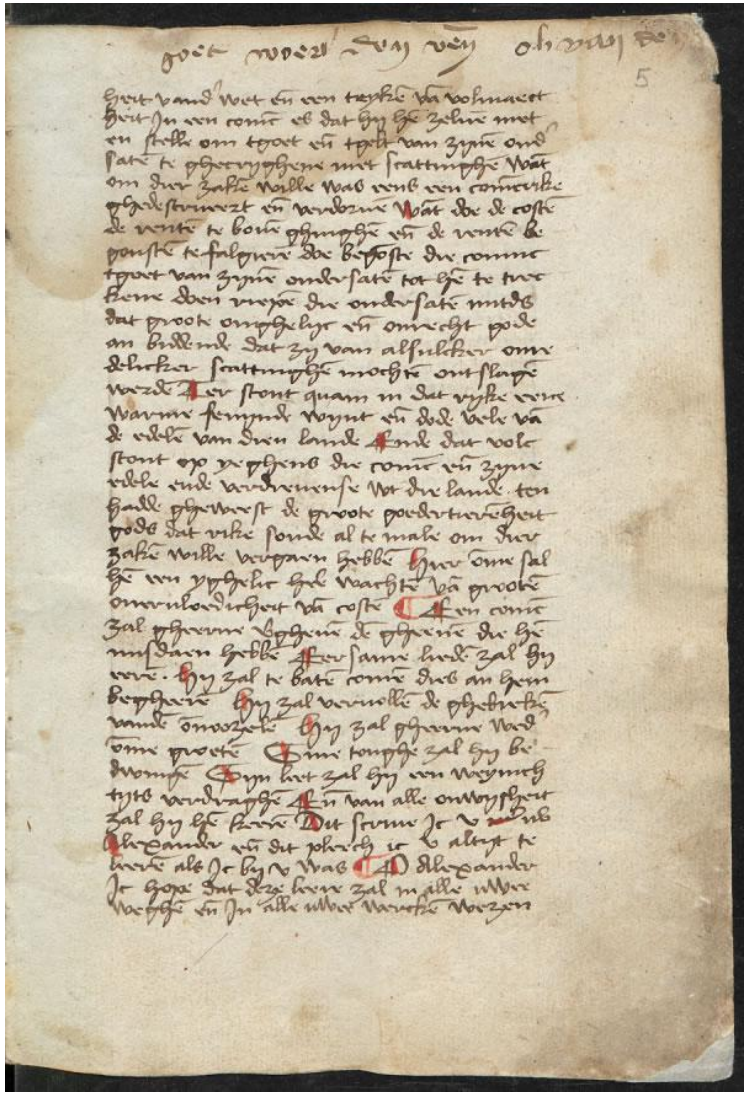
TRANSCRIPTORIUM aims to develop

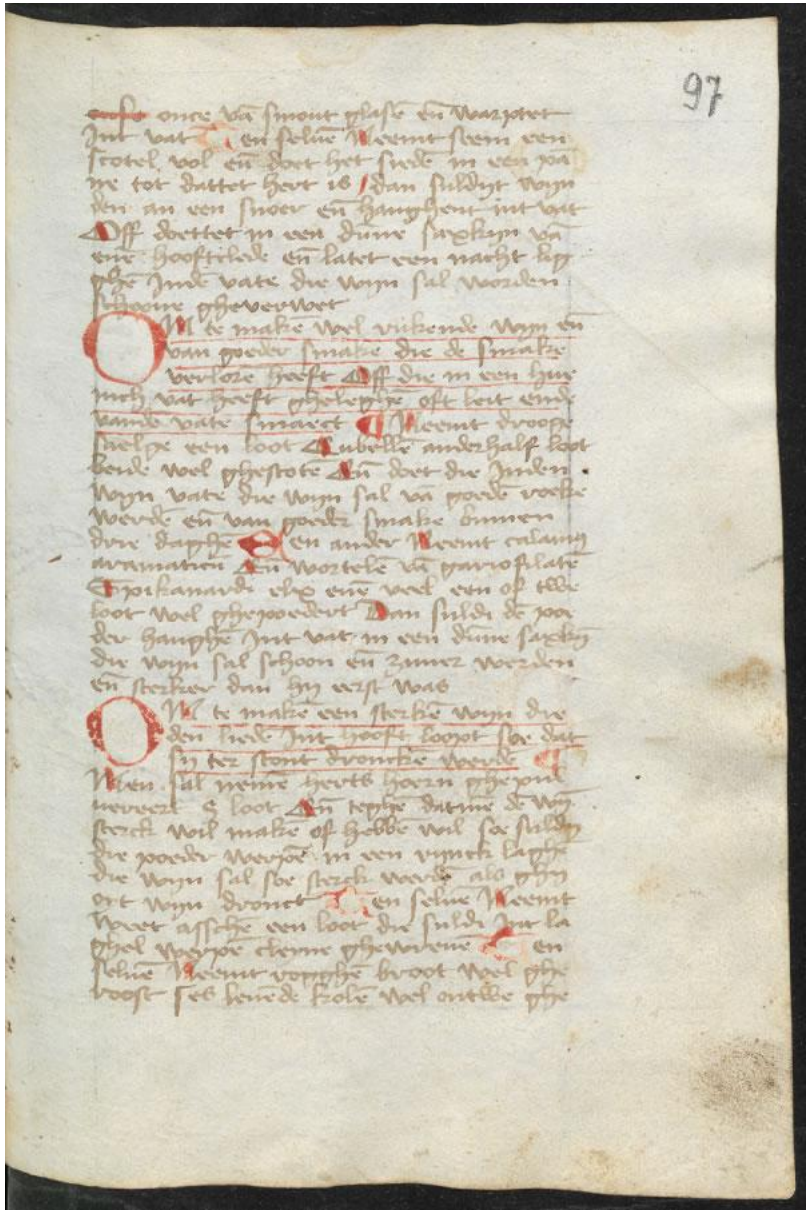
- Solutions for full transcription, indexing and search of historical handwritten document images.
- Using modern, holistic Handwritten Text Recognition (HTR) technology.



18 ghecalcineert
19 ¶ Colorijs
20 ¶ Omme gout te verwen Neemt
21 de 2 deel spaensche groene eñ een
22 deel sal armoniac eñ aysijn ende
23 daer in dan doet opt vier eñ ma-
24 ket heet dan doet af hets ghedaen
25 ¶ Neemt vijf werf alsoe vele ven9
26 als luna eñ smelt hu ven9 eerst
27 eñ dan doeter hu lune in oec ghe-
28 smoltē / of doet het daer in onghe#
29 smolten Eñ dan doetter in arseni-
30 com ghepoedert eñ laet driuen
31 tot schoone wert eñ spieghelt het

Data (Hattem)





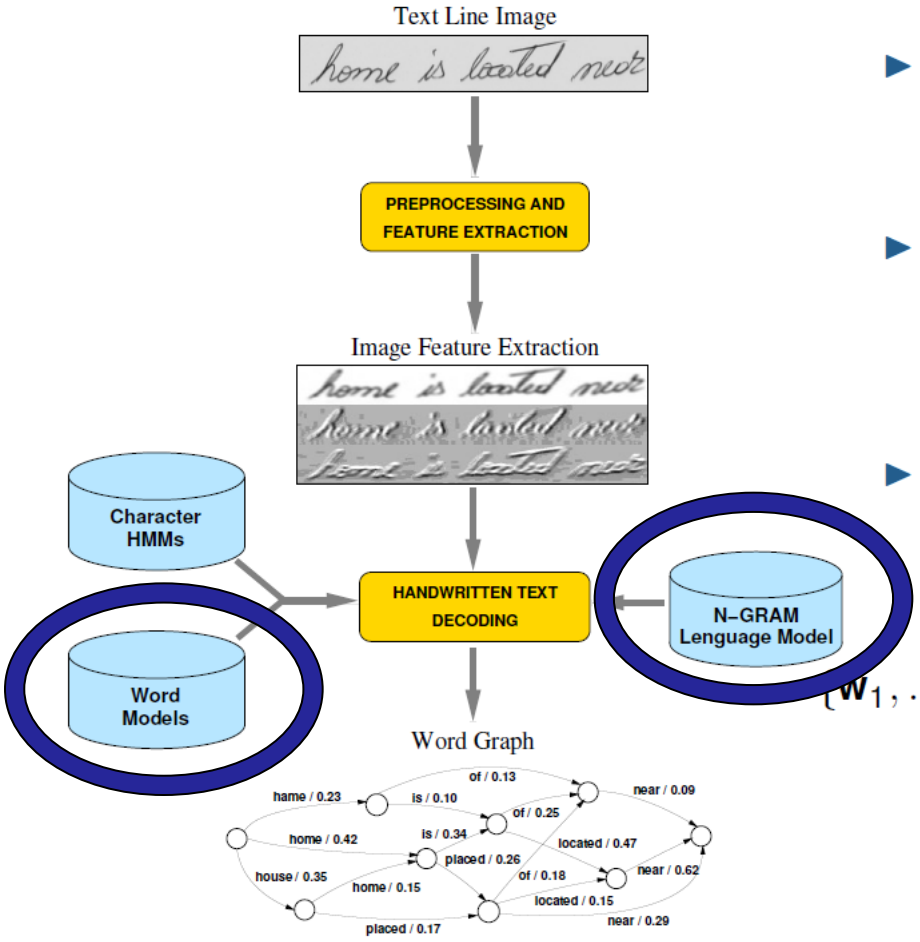
Handwritten text recognition

- Holistic, segmentation-free HTR technology employed here borrows concepts and methods for the field of ASR based on HMMs and N-Grams Language model.
- In contrast with OCR, it does not need any kind of character or word segmentation.

  A. H. Toselli and et. al. Integrated Handwriting Recog- nition and Interpretation using Finite-State Models. Int. Journal of Pattern Recognition and Artificial Intelligence, 18(4):519539, June 2004.
- This HTR technology takes as input pre-processed text line images (without segmenting them into words/characters), and as output produces sequences of recognized words.
- As a byproduct, the HTR process is able to produce a list of n-best recognized hypotheses, which can be embedded into word graphs or lattices.
- The word graph is a fundamental tool not only in HTR, ASR and MT. In this case, word graphs will be used for Interactive Techniques for HTR (T5.3) and for *Key Word Spotting* (KWS) and indexing (T3.4).

Handwritten text recognition

Segmentation-free HTR Approach: Process Scheme:



- **Preprocessing Module**: performing the handwriting style attribute normalization.
- **Feature Extraction Module**: transforming each line image into a sequence of feature vectors: $\mathbf{x} = x_1, x_2, \ldots, x_n,\ x_i \in \Re^D$.
- **Decoding Module**: find a most likely n-best word sequences, $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$, for a given handwritten text line $\mathbf{x}$, according to:

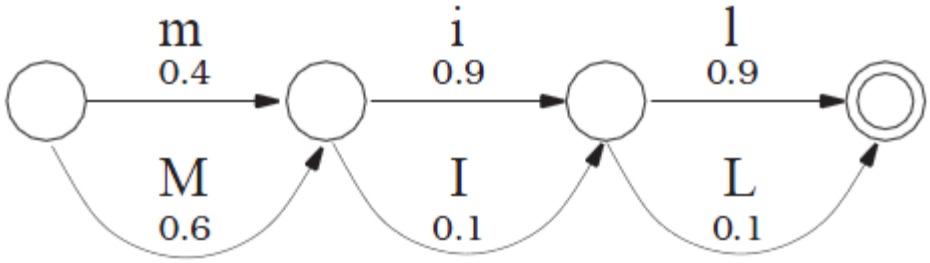
$$\{\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_n\} = \underset{\mathbf{w}}{\text{n-best}}\, P_{\text{HMM}}(\mathbf{x} \mid \mathbf{w}) \cdot P_{\text{N-Gram}}(\mathbf{w})$$

Set of *n*-best hypotheses can conveniently arranged into a so-called: **Word-Graph**.

Linguistic resources for HTR

Two types of linguistic resource needed:

- Lexical models



- N-gram models