Joint Recognition of Handwritten Text and Named Entities with a Neural End-to-end Model

Manuel Carbonell*†, Mauricio Villegas*, Alicia Fornés† and Josep Lladós†
* *omni:us*
Berlin, Germany,
{*manuel,mauricio*}*@omnius.com*
† *Computer Vision Center - Computer Science Department*
Universitat Autonoma de Barcelona, Spain
{*afornes,josep*}*@cvc.uab.es*

***Abstract*—When extracting information from handwritten documents, text transcription and named entity recognition are usually faced as separate subsequent tasks. This has the disadvantage that errors in the first module affect heavily the performance of the second module. In this work we propose to do both tasks jointly, using a single neural network with a common architecture used for plain text recognition. Experimentally, the work has been tested on a collection of historical marriage records. Results of experiments are presented to show the effect on the performance for different configurations: different ways of encoding the information, doing or not transfer learning and processing at text line or multi-line region level. The results are comparable to state of the art reported in the ICDAR 2017 Information Extraction competition, even though the proposed technique does not use any dictionaries, language modeling or post processing.**

***Keywords*-Named entity recognition; handwritten text recognition; neural networks**

I. Introduction

Extracting information from historical handwritten text documents in an optimal and efficient way is still a challenge to solve, since text in these kind of documents are not as simple to read as printed characters or modern handwritten calligraphies [1], [2]. Historical manuscripts contain information that gives an interpretation of the past of societies. Systems designed to search and retrieve information from historical documents must go beyond literal transcription of sources. Indeed it is necessary to shorten the semantic gap and get semantic meaning from the contents, thus the extraction of the relevant information carried out by named entities (e.g. names of persons, organizations, locations, dates, quantities, monetary values, etc.) is a key component of such systems. Semantic annotation of documents, and in particular automatic named entity recognition is neither a perfectly solved problem [3].

Many existing solutions make use of Artificial Neural Networks (ANNs) to transcribe handwritten text lines and then parse the transcribed text with a Named Entity Recognition model, but the precision of those existing solutions is still to improve [1], [2], [4]. One possible approach is to start with already segmented words, by an automatic or manual process, and predict the semantic category using visual descriptors (c.f. [5]) which has the benefit that when the name entity prediction is correct, the transcription would be much easier to predict correctly since it restricts the language model within the corresponding category. The downside is that we rarely have large amounts of word level segmented data, a key for most ANNs proper performance. In case that automatic word segmentation is needed, the whole information extraction process involves three steps which will probably accumulate errors in each of them. Another and most common option is to perform handwritten text recognition (HTR) first and then named entity recognition (NER). An advantage of this approach is that it has one less step than the previous explained approach, but it has the counterpart that if the transcription is wrong, the NER part is affected.

Recent work in ANNs suggests that using models that solve tasks as general as possible, might give similar or better performance than concatenating subprocesses due to error propagation in the different steps, as shown in [6], [7]. This is the main motivation of this work, and consequently we propose a single convolutional-sequential model to jointly perform transcription and semantic annotation. Adding a language model, the transcription can be restricted to each semantic category and therefore improved. The contribution of this work is to show the improvement when joining a sequence of processes in a single one, and thus, avoiding to commit accumulation of errors and achieving generalization to emulate human-like intelligence.

Some examples of historical handwritten text documents include birth, marriage and defunction records which provide very meaningful information to reconstruct genealogical trees and track locations of family ancestors, as well as give interesting macro-indicators to scholars in social sciences and humanities. The interpretation of such types of documents unavoidably requires the identification of named entities. As experimental scenario we illustrate the performance of the proposed method on a collection of handwritten marriage records.

The rest of the paper is organized as follows: Next section explains the task being considered. In section III we review the state of the art work in HTR and NER. In IV we explain our model architecture, ground truth setup and training details. In Section V we analyze the results for the different configurations and last in VI we give the conclusions.

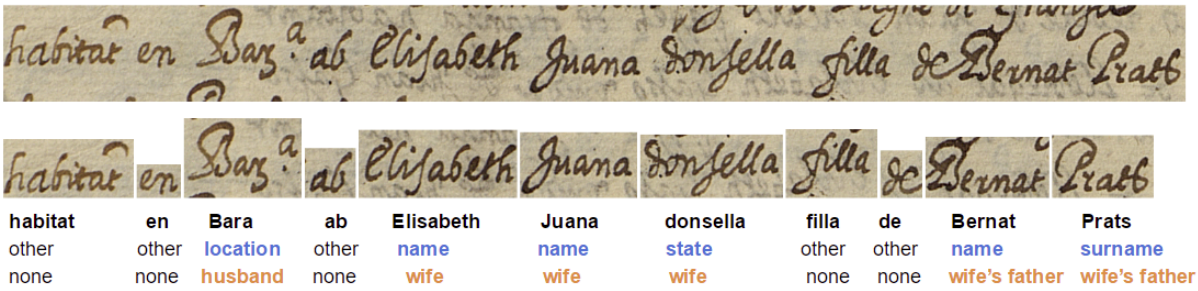


Figure 1. An example of a document line annotation from [4].

Table I: Semantic and person categories in the IEHHR competition

Semantic	**Person**
Name	Wife
Surname	Husband
Occupation	Wife's father
Location	Wife's Mother
Civil State	Husband's father
Other	Husband's mother
	Other person
	None

II. The Task: Information Extraction in Marriage Records

The approach presented in this paper is general enough to be applied to many information extraction tasks, but due to time constraints and our access to a particular dataset, the approach is evaluated on the task of information extraction in a system for the analysis of population records, in particular handwritten marriage records. It consists of transcribing the text and to assign to each word a semantic and person category, i.e. to know which kind of word has been transcribed (name, surname, location, etc.) and to what person it refers to. The dataset and evaluation protocol are exactly the same as the one proposed in the ICDAR 2017 Information Extraction from Historical Handwritten Records (IEHHR) competition [4]. The semantic and person categories to identify in the IEHHR competition are listed in table I.

Two tracks were proposed. In the basic track the goal is to assign the semantic class to each word, whereas in the complete track it is also necessary to identify the person. An example of both tracks is shown in Figure 1.

The dataset for this competition contains 125 pages with 1221 marriage records (paragraphs), where each record contains several text lines giving information of the wife, husband and their parents' names, occupations, locations and civil states. The text images are provided at word and line level, naturally having the increased difficulty of word segmentation when choosing to work with line images. More details of the dataset can be found in table II.

Table II: Marriage Records dataset distribution

	Train	Validation	Test
Pages	90	10	25
Records	872	96	253
Lines	2759	311	757
Words	28346	3155	8026

Out of vocabulary words: 5.57 %

III. State of the art

Recent work shows that neural models allow generalization of problems that earlier were solved separately [7]. This idea can also be applied to information extraction from handwritten text documents which consists of HTR followed by NER. From the HTR side there is still a long way to improve until human level transcription is achieved [8]. Attention models have helped to understand the inside behavior of neural networks when reading document images but still have lower accuracy than Recurrent Neural Network with Connectionist Temporal Classification (RNN+CTC) approaches [9].

Named entity recognition is the problem of detecting and assigning a category to each word in a text, either at part-of-speech level or in pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. The goal is to select and parse relevant information from the text and relationships within it. One could think that it would be sufficient to keep a list of locations, common names and organizations, but the case is that these lists are rarely complete, or one single name can refer to different kind of entities. Also it is not easy to detect properties of a named entity and how different named entities are related to each other. Most widely used kind of models for this task are *conditional random fields* (CRFs), which were the state of the art technique for some time [10], [11].

In the area of Natural Language Processing, Lample et al. [3] proposed a combination of *Long Short-term Memory networks* (LSTMs) and CRFs, obtaining good results for the CoNLL2003 task. The problem is similar to the one we are facing, except that it starts from raw text. In this work the input to the system are images of handwritten text lines, for which it is not even known how many characters or words are present. This undoubtedly introduces a higher difficulty.

In Adak's work [12] a similar end-to-end approach from image to semantically annotated text is proposed, but in that case the key relies in identifying capital letters to

detect possible named entities. The problem is that in many cases, such as in the IEHHR competition [4] dataset, named entities do not always have capital letters, and also, it is a task-specific approach that could not be used in many other cases.

Finally, another concept that can help to improve the quality of our models' prediction is curriculum learning [13]. Letting the model look at the data in a meaningful and ordered way, such that the difficulty of prediction goes from easy to hard, and therefore, can make the training evolve with a much better performance.

IV. Methodology

The main goal of this work is to explore a few possibilities for a single end-to-end trainable ANN model that receives as input text images and gives as output transcripts, already labeled with their corresponding semantic information. One possibility to solve it could be to propose a ANN with two sequence outputs, one for the transcript and the other for the semantic labels. However, keeping an alignment between these two independent outputs complicates a solution. An alternative would be to have a single sequence output that combines the transcript and semantic information, which is the approach taken here. There are several ways in which this information can be encoded such that a model learns to predict it. The next subsection describes the different ways of encoding it that were tested in this work. Then there are subsections describing the architecture chosen for the neural network, the image input and characteristics of the learning.

A. Semantic encoding

The first variable which we explored is the way in which ground truth transcript and semantic labels are encoded so that the model learns to predict them. To allow the model to recognize words not observed during training (out-of-vocabulary) the symbols that the model learns are the individual characters and a space to identify separation between words. For the semantic labels special tags are added to the list of symbols for the recognizer. The different possibilities are explained below.

1) Open & close separate tags: In the first approach, the words are enclosed between **opening and closing tags** that encode the semantic information. Both the category and the person have independent tags. Thus, each word is encoded by starting with opening category and person symbols, followed by a symbol for each character and ends by closing person and category symbols. The "other" and "none" semantics are not encoded. For example, the ground truth of the image shown in Figure 1 would be encoded as:

```
h a b i t a t {space} e n {space}
<location> <husband> B a r a </husband>
</location> {space} a b {space} <name>
<wife> E l i s a b e t h </wife>
</name> ...
```

This kind of encoding is not expected to perform well in the IEHHR task, since tags are assigned to only one word at a time, so it is redundant to have two tags for each word. However, in other tasks it could make sense having opening and closing tags and this is why it has been considered in this work.

2) Single separate tags: Similar to the previous approach, in this case both category and person tags are independent symbols but there is only one for each word added before the word. Thus, the ground truth of the previous example would be encoded as:

```
h a b i t a t {space} e n {space}
<location/> <husband/> B a r a {space}
a b {space} <name/> <wife/> E l i s a b
e t h {space} J u a n a {space}
<state/> <wife/> {space} d o n s e l l
a ...
```

3) Change of person tag: In this variation of the semantic encoding the person label is only given if there is a **change of person**, i.e. the person label indicates that all the upcoming words refer to that person until another person label comes, in contrast to previous approaches where we give the person label for each word. This approach is possible due to the structured form of the sentences in the dataset. As we can see in Figure 2 the marriage records give the information of all the family members without mixing them.

```
<wife/> <name/> E l i s a b e t h
{space} <name/> J u a n a {space}
<state/> d o n s e l l a ...
```

4) Single combined tags: The final possibility tested for encoding the named entity information is to **combine category and person** labels into a single tag. So the example would be as:

```
h a b i t a t {space} e n {space}
<location_husband/> B a r a {space} a b
{space} <name_wife/> E l i s a b e t h
{space} <name_wife/> J u a n a {space}
<state_wife/> d o n s e l l a ...
```

B. Level of input images: lines or records

The IEHHR competition dataset includes manually segmented images at word level. But to lower ground truthing cost or avoid needing a word segmentator, we will assume that only images at line level are available. Having text line images then the obvious approach is to give the system individual line images for recognition. However, there are semantic labels that would be very difficult to predict if only a single line image is observed due to lack of context. For example, it might be hard to know if the name of a person corresponds to the husband or the father of the wife if the full record is not given. Because of this, in the experiments we have explored having as input both text line images and full marriage record images, concatenating all the lines of a record one after the other.

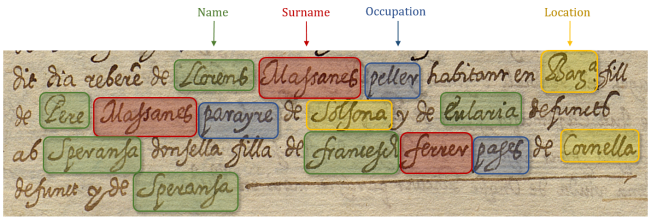


Figure 2. Reading the whole record makes it easier to transcribe as well as to identify the semantic categories based on context information.

C. Transfer learning

The next variable we examined was the effect of the use of **transfer learning** from a previously trained model for HTR. Transfer Learning consists of training for the same or a similar task (HTR) using other datasets, and then fine tune it for our purpose, in our case HTR+NER. To perform transfer learning from a generic HTR model, the softmax layer is removed and replaced with a softmax that allows as an output the activations for the number of possible classes in the fine tuning step. In our case, they will be all the characters in the alphabet plus the semantic labels. In the experiments for transfer learning we have tested only one HTR model that was trained with the following datasets: IAM [14], Bentham [15], Bozen [16], and some datasets used by us internally: IntoThePast, Wiensanktulrich, Wienvotivkirche and ITS.

D. Curriculum Learning

The last variation that we propose is curriculum learning i.e. start with easier demands to the model and then increase the difficulty. In this case this method can be interpreted as starting by learning to transcribe single text lines, and when the training is finished, continue with learning to transcribe images of a whole marriage record.

E. Model architecture and training

In this work we use a CNN+BLSTM+CTC model, which is one of the most common models for performing HTR exclusively, although other HTR models could be used as well. In particular, the architecture consists of 4 convolutional layers with max pooling followed by 3 stacked BLSTM layers. The detailed model architecture is shown in Figure 3.

To train the model we use the Laia HTR toolkit [17] which uses Baidu's parallel CTC [18] implementation, which consists of minimizing the loss or "objective" function

$$O^{ML}(S, \mathcal{N}_w) = -\sum_{(x,z)\in S} ln(p(z|x)) \qquad (1)$$

where S is the training set, x is the input sequence (visual features), z is the sequence labeling (transcription) for x and

$$\mathcal{N}_w : (\mathbb{R}^m)^T \mapsto (\mathbb{R}^n)^T \qquad (2)$$

is a recurrent neural network with m inputs, n outputs and weight vector w. The probabilities of a labeling of an input sequence are calculated with a dynamic programming algorithm called "forward-backward".

Some special features of our model are that the activation function for the convolutional layers is leaky ReLu $f(x) = x$ if $x > 0.01,\ 0.01x$ otherwise.

We also use batch normalization to reduce *internal covariate shift* [19].

V. Results

We compare the performance of our methods[1] with the results of the participants of the IEHHR competition in [4] thereby using the same metric, see Table III. The evaluation metric counts the words that were correctly transcribed and annotated with their category and person label with respect to the total amount of words in the ground truth. For those words that were not correctly transcribed but the category and person labels match one or more words in the ground truth, we add to the score 1 - CER (character error rate) on the best matching word. This means that the named entity recognition part is vital for a good score, since a perfect transcription will count as 0 in the score if its named entity is incorrectly detected.

We can observe in the results that our best performance is reached when receiving the whole marriage record, which is probably due to the help of contextual information. For example, it can benefit the detection of named entities composed of several words when they are written in separate consecutive lines. Also we observe that the best performing encoding of the semantic labels is the combined tags setup. This can be due to the lower amount of symbols to predict, which might require to store less long term dependencies in the network.

The most significant improvement was achieved when picking our best performing configuration and running it with an alternative line extraction. In the competition, the text lines were extracted by including all the bounding boxes of the words within every line. As a result, when there are large ascenders and descenders, the bounding box of the line is too wide, including sections of other text lines. In order to cope with this limitation, we used the XML containing the exact location of the segmented words within a page, and for the y-coordinates, we used a weighted (by the words widths) average of upper and lower limits of the word bounding boxes. As expected, the performance highly improves because the segmentation of the text lines is more accurate. However, this result is not directly comparable to the other participants's methods because the segmentation is different.

In Figure 4 we show some examples of committed errors. We can see that they consist of small typos that are understandable when looking at the text images. It is definitely difficult to transcribe certain names that have never been seen before. The proposed approach could be combined with a category-based language model [1] which could potentially improve the results.

[1]Scripts used for the experiments available at http://doi.org/10.5281/zenodo.1174113

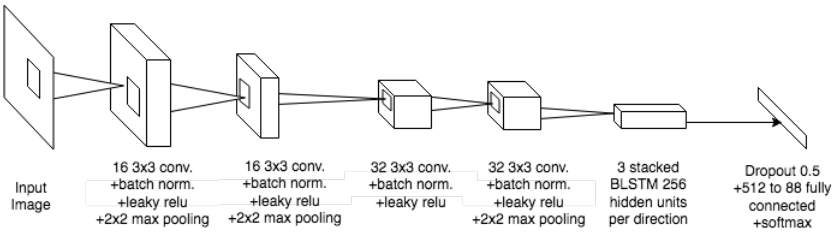
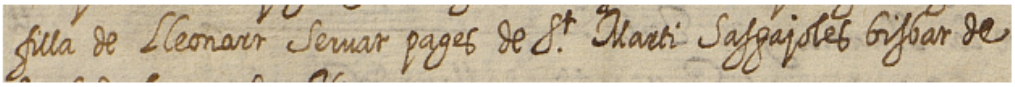
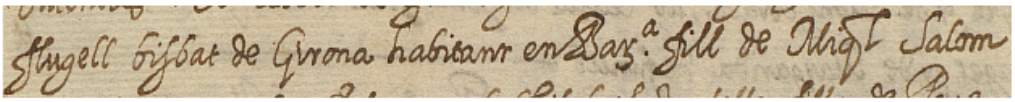


Figure 3. Used model architecture



Predicted: filla de Lleonart Servat pages de **St Marti Sassajoles bisbat de Vich**
Original: filla de Lleonart Servat pages de **St Marti Sasgajoles bisbat de Vich**



Predicted: **Argell bisbat de Grona** habitant en Bara fill de Miquel Salom
Original: **flugell bisbat de Girona** habitant en Bara fill de Miquel Salom

Figure 4. Some of the errors committed in the predictions

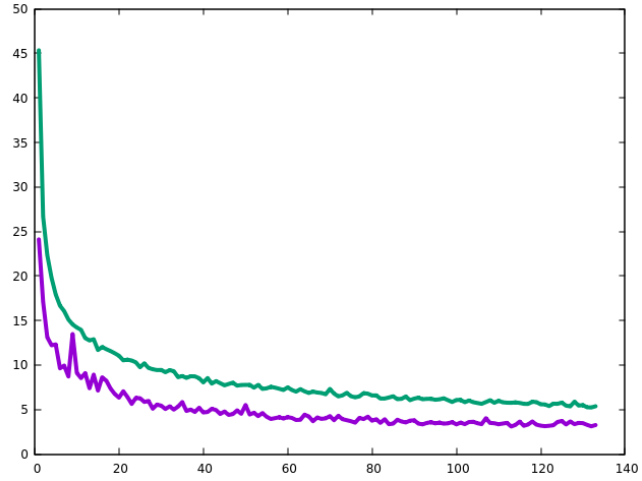


Figure 5. Train and validation (green and violet respectively) CER (%).

Our best performing model took 4 hours 38 to run 133 training epochs with a NVIDIA GTX 1080 GPU. The train and validation error rates can be seen in Figure 5. As training configuration we used an adversarial regularizer [20] with weight 0.5, an initial learning rate of $5 \cdot 10^{-4}$ with decay factor of 0.99 per epoch and batch size 6.

VI. Conclusion

In this paper we have proposed to solve a complex task (i.e. text recognition and named entity recognition) with a single end-to-end neural model. Our first conclusion is that, also in information extraction problems, a generic model for solving two subsequent tasks can perform at least similarly as two separated models. This is true even if there is less prepared data (record level images instead of a sequence of word images) and we do not make use of task specific tools like dictionaries or language model.

By investigating different ways of encoding the image transcripts and semantic labels we have shown that the recognition performance is highly affected, even though it is indeed representing the same information. Also, curriculum learning (first text lines and then records) can make the model reach a higher final prediction accuracy.

Future work would include the use of language models to improve the accuracy of the predictions, the effect of automatic text line and record detection, and also, to evaluate our method in other datasets.

Acknowledgments

This work has been partially supported by the Spanish project TIN2015-70924-C2-2-R, the grant 2016-DI-095 from the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya, the Ramon y Cajal Fellowship RYC-2014-16831, the CERCA Programme /Generalitat de Catalunya, and RecerCaixa (XARXES, 2016ACUP-00008), a research program from Obra Social "La Caixa" with the collaboration of the ACUP.

References

[1] V. Romero, A. Fornes, E. Vidal, and J. A. Sanchez, "Using the mggi methodology for category-based language modeling in handwritten marriage licenses books," in *15th international conference on Frontiers in Handwriting Recognition*, 2016.

[2] A. H. Toselli, E. Vidal, V. Romero, and V. Frinken, "Hmm word graph based keyword spotting in handwritten document images," *Inf. Sci.*, vol. 370, no. C, pp. 497–518, Nov. 2016. [Online]. Available: https://doi.org/10.1016/j.ins.2016.07.063

Table III: Average scores of the experiments compared with the IEHHR competition participants' methods.

Method	Segm. Level	Proc. Level	Track Basic	Track Complete
IEHHR competition results				
Hitsz-ICRC-1 CNN HTR+NER	Word	Record*	87.56	85.72
Hitsz-ICRC-2 ResNet HTR+NER	Word	Record*	94.16	91.97
Baseline HMM+MGGI	Line	Record	80.24	63.08
CITlab-ARGUS-1 LSTM+CTC+regex	Line	Record†	89.53	89.16
CITlab-ARGUS-2 LSTM+CTC +OOV+regex	Line	Record†	**91.93**	**91.56**
Results of our experiments				
Separate-single tags	Line	Line	73.49	61.96
Separate- open-close tags	Line	Line	73.70	64.09
Combined-single tags	Line	Line	87.96	80.74
Combined-single tags + transfer learn	Line	Line	87.01	80.05
Change person tag + transfer learn	Line	Record	84.41	80.51
Combined-single tags + transfer learn	Line	Record	86.58	84.72
Combined-single tags + transfer learn + curriculum learn	Line	Record	**90.58**	**89.39**
Combined-single tags + transfer learn + curriculum learn + alt. line extraction	Word‡	Record‡	**96.39**‡	**96.63**‡

*HTR is word based.

†Posterior character probabilities computed at line level.

‡Not fair to compare with the IEHHR results because it uses a different segmentation (alternative line extraction) than the one provided in the competition.

[3] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *CoRR*, vol. abs/1603.01360, 2016. [Online]. Available: http://arxiv.org/abs/1603.01360

[4] A. Fornés, V. Romero, A. Baró, J. I. Toledo, J. A. Sanchez, E. Vidal, and J. Lladós, "Competition on information extraction in historical handwritten records," in *International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2017.

[5] J. I. Toledo, S. Sudholt, A. Fornés, J. Cucurull, G. A. Fink, and J. Lladós, *Handwritten Word Image Categorization with Convolutional Neural Networks and Spatial Pyramid Pooling*. Cham: Springer International Publishing, 2016, pp. 543–552.

[6] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *CoRR*, vol. abs/1606.04404, 2016. [Online]. Available: http://arxiv.org/abs/1606.04404

[7] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," *CoRR*, vol. abs/1604.07316, 2016. [Online]. Available: http://arxiv.org/abs/1604.07316

[8] T. Bluche, "Joint Line Segmentation and Transcription for End-to-End Handwritten Paragraph Recognition," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 838–846.

[9] T. Bluche, J. Louradour, and R. O. Messina, "Scan, attend and read: End-to-end handwritten paragraph recognition with MDLSTM attention," *CoRR*, vol. abs/1604.03286, 2016. [Online]. Available: http://arxiv.org/abs/1604.03286

[10] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. [Online]. Available: http://dl.acm.org/citation.cfm?id=645530.655813

[11] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 363–370. [Online]. Available: https://doi.org/10.3115/1219840.1219885

[12] C. Adak, B. B. Chaudhuri, and M. Blumenstein, "Named entity recognition from unstructured handwritten document images," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, April 2016, pp. 375–380.

[13] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 41–48. [Online]. Available: http://doi.acm.org/10.1145/1553374.1553380

[14] U. v. Marti and H. Bunke, "A full english sentence database for off-line handwriting recognition," in *In Proc. Int. Conf. on Document Analysis and Recognition*, 1999, pp. 705–708.

[15] J. A. Sánchez, V. Romero, A. H. Toselli, and E. Vidal, "Icfhr2014 competition on handwritten text recognition on transcriptorium datasets (htrts)," in *2014 14th International Conference on Frontiers in Handwriting Recognition*, Sept 2014, pp. 785–790.

[16] J. Sánchez, V. Romero, A. Toselli, and E. Vidal, "ICFHR2016 competition on handwritten text recognition on the READ dataset," in *ICFHR*. IEEE, 2016, pp. 630–635.

[17] J. Puigcerver, D. Martin-Albo, and M. Villegas, "Laia: A deep learning toolkit for htr." GitHub, 2016, gitHub repository.

[18] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 369–376. [Online]. Available: http://doi.acm.org/10.1145/1143844.1143891

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[20] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014.