Normalized vs Diplomatic Annotation: A Case Study of Automatic Information Extraction from Handwritten Uruguayan Birth Certificates

Natalia Bottaioli[1,2,3][0009-0005-3606-908X], Solène Tarride[4][0000-0001-6174-9865], Jérémy Anger[1][0009-0007-3319-6037], Seginus Mowlavi[1][0009-0005-7774-9217], Marina Gardella[5][0000-0003-2465-2014], Antoine Tadros[1][0000-0002-7885-903X], Gabriele Facciolo[1][0000-0002-8855-8513], Rafael Grompone von Gioi[1][0000-0002-6309-7116], Christopher Kermorvant[4][0000-0002-7508-4080], Jean-Michel Morel[6][0000-0002-6108-897X], and Javier Preciozzi[2,3][0000-0002-9287-4826]

[1] Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, France
[2] Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay
[3] Digital Sense, Montevideo, Uruguay
[4] TEKLIA, Paris, France
[5] IMPA, Rio de Janeiro, Brasil
[6] City University of Hong Kong, Hong Kong

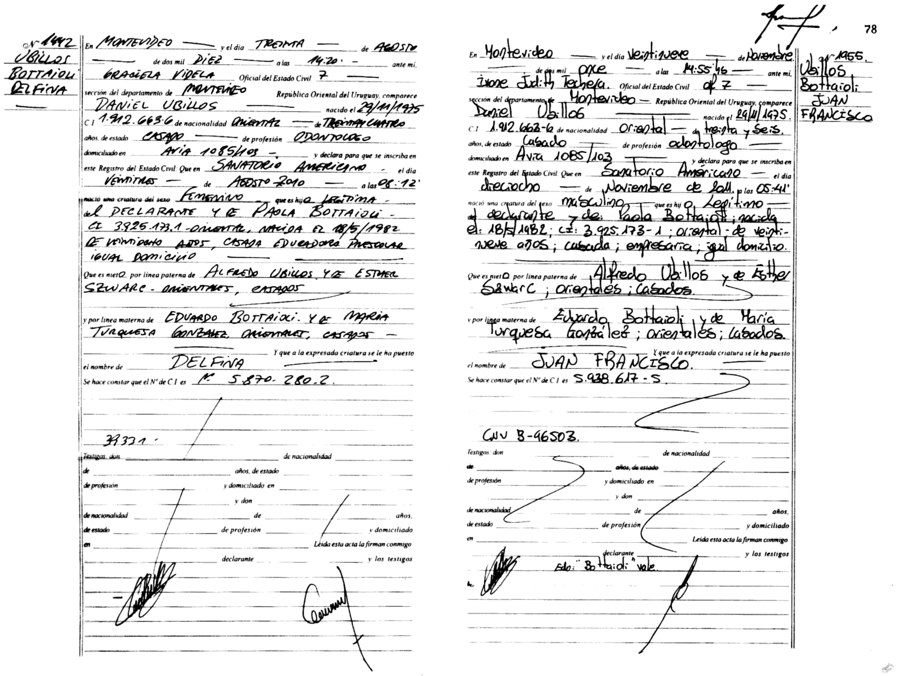
Abstract. This study evaluates the recently proposed Document Attention Network (DAN) for extracting key-value information from Uruguayan birth certificates, handwritten in Spanish. We investigate two annotation strategies for automatically transcribing handwritten documents, fine-tuning DAN with minimal training data and annotation effort. Experiments were conducted on two datasets containing the same images (201 scans of birth certificates written by more than 15 different writers) but with different annotation methods. Our findings indicate that normalized annotation is more effective for fields that can be standardized, such as dates and places of birth, whereas diplomatic annotation performs much better for fields containing names and surnames, which can not be standardized.

Keywords: Automatic information extraction · Handwritten text recognition · Birth certificates transcription · Normalized and diplomatic annotation

1 Introduction

Civil Registration and Vital Statistics (CRVS) is defined by the United Nations as the "continuous, permanent, compulsory and universal recording of the occurrence and characteristics of vital events of the population in accordance with the law".[1] CRVS is strongly related to article 6 of the Universal Declaration of

[1] https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Handbooks/crvs/crvs-mgt-E.pdf



N° 1442
UBILLOS
BOTTAIOLI
DELFINA

En MONTEVIDEO y el dia TREINTA de AGOSTO
de dos mil DIEZ a las 14:20. ante mi,
GRACIELA VIDELA Oficial del Estado Civil 7
sección del departamento de MONTEVIDEO República Oriental del Uruguay, comparece
DANIEL UBILLOS nacido el 29/11/1975
C.I. 1.912.663-6 de nacionalidad ORIENTAL de TREINTA Y CUATRO
años, de estado CASADO de profesión ODONTOLOGO
domiciliado en AVIA 1085/103 - y declara para que se inscriba en
este Registro del Estado Civil: Que en SANATORIO AMERICANO - el dia
VEINTITRES de AGOSTO 2010 a las 08:12'
nació una criatura del sexo FEMENINO que es hija LEGITIMA -
del DECLARANTE Y DE PAOLA BOTTAIOLI -
CI 3.925.173.1 - ORIENTAL, NACIDA EL 18/5/1982
DE VEINTIOCHO AÑOS, CASADA EDUCADORA PREESCOLAR
IGUAL DOMICILIO
Que es nieta por línea paterna de ALFREDO UBILLOS Y DE ESTHER
SZWARC - ORIENTALES, CASADOS
y por línea materna de EDUARDO BOTTAIOLI. Y DE MARIA
TURQUESA GONZALEZ. ORIENTALES, CASADOS -
Y que a la expresada criatura se le ha puesto
el nombre de DELFINA
Se hace constar que el N° de C.I. es N° 5.870-280.2.
39331.
Testigos: don de nacionalidad
de años, de estado
de profesión y domiciliado en
y don
de nacionalidad de años,
de estado de profesión y domiciliado
en Leida esta acta la firman conmigo
declarante y los testigos

(a) Left margin

78

En Montevideo y el dia Veintinueve de Noviembre
de dos mil once a las 14:55'46 ante mi,
Ivone Judith Techera. Oficial del Estado Civil of 7
sección del departamento de Montevideo República Oriental del Uruguay, comparece
Daniel Ubillos nacido el 29/11/1975.
C.I. 1.912.663-6 de nacionalidad Oriental de treinta y seis.
años, de estado Casado de profesión odontologo
domiciliado en Avia 1085/103 y declara para que se inscriba en
este Registro del Estado Civil: Que en Sanatorio Américano el dia
dieciocho de Noviembre de 2011. a las 05:41'
nació una criatura del sexo masculino que es hijo Legitimo
del declarante y de: Paola Bottaioli; nacida
el: 18/5/1982; CI: 3.925.173-1; oriental - de veintinueve años; casada; empresaria; igual domicilio.
Que es nieto por línea paterna de Alfredo Ubillos y de Esther
Szwarc; orientales; Casados.
y por línea materna de Eduardo Bottaioli y de María
Turquesa González; orientales; casados.
Y que a la expresada criatura se le ha puesto
el nombre de JUAN FRANCISCO.
Se hace constar que el N° de C.I. es 5.938.617-5.
CNV B-96503.
Testigos: don de nacionalidad
de años, de estado
de profesión y domiciliado en
y don
de nacionalidad de años,
de estado de profesión y domiciliado
en Leida esta acta la firman conmigo
declarante y los testigos
Edo: "Bottaioli" vale.

N° 1955.
Ubillos
Bottaioli
JUAN
FRANCISCO

(b) Right margin

Fig. 1: Two scanned Uruguayan birth certificates, one with the margin on the left and the other one on the right.

Human Rights, which states that everybody has the right to a legal identity.[2] Having a birth certificate gives to individuals legal evidence of their name, place and date of birth, and parents' names (among other data). In most countries, CRVS are key to ensure the right to have a legal identity, and this is clearly reflected by the fact that 67 of the 232 Sustainable Development Goals[3] are related to CRVS.[4] Yet, more than 1 billion people worldwide do not have a document that serves as a proof of legal identity.

Because most CRVS were created long before the advent of computers, the majority of their registries are paper-based handwritten documents. This huge amount of documents is only recently being scanned and integrated into digital systems. Belize, Jamaica and Perú are some of the countries in America that are nowadays in the process of scanning historical CRVS data.

Here, we will focus on the case of Uruguay, where CRVS responsibility is held by a National Civil Registry Office, named the *Dirección General del Reg-*

[2] https://www.undp.org/africa/blog/having-legal-identity-fundamental-human-rights
[3] https://sdgs.un.org/es/goals
[4] https://getinthepicture.org

istro de Estado Civil.[5] Initially named *Registro de Estado Civil*, this office exists since 1879, when the government started keeping records of births, marriages, recognitions and deaths nationwide. In 2012, Uruguay's capital city, Montevideo, started registering birth certificates digitally, a process that was completed for the rest of the country only by 2022. Figure 1 shows two scanned birth certificates from Montevideo, one in which the margin is on the right and another one where the margin is on the left. It is worth mentioning that, due to the long period in which registries were generated manually, several different templates exist, and not all formats contain the same information. Even when the format is the same, the way of filling the data is not homogeneous in time.

A digitization process has been gradually taking place in some departments but not yet nationwide: books have been scanned and, therefore, digital images of vital events have been generated. In order to make scanned images searchable and linked to other relevant registries, this digitization process is generally followed by a manual transcription of, at least, part of the data. This is a very demanding and time-consuming task, needless to say prone to errors. It is in this context that automatically processing these registries becomes of great interest.

The process of automatically recognizing the text present in an image requires both the interpretation of handwritten text (known as handwritten text recognition) and the layout interpretation of the scanned document (known as document layout analysis). This process generally requires the adaptation of the existing methods to a specific context. In [11], authors propose a Document Attention Network (DAN) for handwritten document recognition, where the adaptation process does not require to manually segment the different fields present in an image. In fact, only the scanned images and the fields to be transcribed with their corresponding annotations are required.

The objective of this work is twofold. First, we want to understand whether DAN can be adapted to work on the specific problem of handwritten document recognition for Uruguayan birth certificates, using very few data for the adaptation process. This analysis is relevant since DAN was trained for a different context (letters) and a different language (French). Second, we want to analyze two different annotation strategies. The first one consists of using **normalized** annotations, directly obtained from the Civil Registry Office. The second one uses **diplomatic** transcriptions (i.e. those that transcribe each and every character, respecting upper and lower cases, accents, etc). It is worth noting that in the first strategy, the data used to train the model was neither designed nor generated to train a deep learning model, while in the second one annotations were specifically generated for this task.

The rest of the article is organized as follows: Section 2 describes the state of the art regarding handwritten document processing. Section 3 reviews DAN, which is the model evaluated in this work. The Uruguayan birth certificates dataset is described in Section 4. Our two major experiments are presented in Section 5, while Section 6 discusses the results. Conclusions are drawn in Section 7, where future work is also described.

[5] https://www.gub.uy/ministerio-educacion-cultura/registro-civil

2 Related work

Handwritten document processing aims at producing some desired textual information from image inputs. It, therefore, sits at the intersection of computer vision (CV) and natural language processing (NLP). The difference in the modalities of these two fields has encouraged a sequential approach of the problem. The typical document processing pipeline involves one or several steps which perform transcription, i.e. mapping of the input image to a faithful textual output, followed by a natural language understanding step addressing a downstream task, e.g. key information extraction or named entity recognition. The large improvements brought by deep learning in CV [17] and NLP [19], while strengthening the different steps of the sequential document processing pipeline, have also recently enabled integrated approaches. In the latter, algorithms perform the desired downstream task directly from the image input in an end-to-end fashion.

The two aforementioned approaches, sequential and integrated, lie at both ends of a spectrum that we briefly describe here. The sequential approach decomposes the workflow into a pipeline of simple subtasks, most often text line detection, handwritten text recognition (HTR), and finally the desired natural language understanding task. Current text line detection methods [23, 20] adapt the object detection literature, in particular CNNs [29], to the domain of document images. HTR methods [4, 26] follow an encoder-decoder architecture, making use of decoders popularized in the speech-processing context such as hidden Markov models or recurrent neural networks with connectionist temporal classification [13]. Finally, the NLP step is often carried out by a pre-trained transformer encoder model [12] with simple decoder heads fine-tuned on the given task [3, 27]. Variations on this workflow include fusing and jointly training the text line detection and HTR steps [5, 37], or enriching the NLP step with the additional input modalities of layout and/or image besides text [15, 24, 35]. This added integration can bring specific benefits: for example, the former gains efficiency in the image embedding by using a single image encoder, and the latter augments the performance of the NLP step by allowing it to retain information from the input image or the intermediate text line detection output. More recently, the potential of transformers as decoders of visual embeddings [18] has allowed for fully end-to-end models for document processing [11, 16].

An important part of the literature focuses on applying this body of research to settings arising from an industrial need [22, 34, 28, 9, 32, 6]. This type of work contributes back to the theoretical aspects in two key respects. First, it addresses the need for evaluation protocols able to eliminate the influence of benchmark-specific optimizations [7], without expensive searches of the best training procedure for each method. Second, carrying out the entire document processing workflow enables the discovery of important insights beyond the issues of model design and training. For instance, a segmented approach benefits from the availability of powerful pre-trained algorithms [34] but runs the risk of errors accumulating across the steps [31]. Hence, the choice of each algorithm needs to be guided by its input domain, e.g. the NLP module needs to be robust to “recognition error noise” in the input [2, 21]. Conversely, [34, 25] highlight the influence of the training dataset creation philosophy on the performance of HTR models and, thus, of sequential pipelines. Indeed, not only do annotations need to contain full transcriptions (even if a downstream key information extraction task only concerns a subset of that information), but transcriptions also need to stay faithful to the text (this is known as "diplomatic annotation"). Diplomatic annotation becomes an issue when many database creators apply standardization to dates and names, expand abbreviations or introduce them, sometimes even without following consistent guidelines [8]. By contrast, integrated approaches can deal with partial annotation of training data [31]. In addition, it has been shown that their multi-modality yields an increase in language modeling capabilities [33], which could translate into better flexibility with respect to unfaithfully annotated datasets.

3 Model description

For our experiments, we use DAN [11, 1]. We chose it for its off-the-shelf ability to be fine-tuned on data with a minimal amount of annotation effort, as it does not require bounding box-like annotation of text in the full image. Being an end-to-end architecture, it also allows us to study the influence of different annotation strategies in the context of an integrated approach to document processing.

The architecture of DAN follows the encoder-decoder design. The encoder is the fully convolutional network described in [10], with the fixed 2D positional encoding of [30]. The decoder follows the original design of the auto-regressive transformer decoder [36], including the fixed 1D positional encoding. It generates tokens from a dataset-dependent dictionary $\mathcal{D} = \mathcal{A} \cup \mathcal{S} \cup \{\langle \text{eot} \rangle\}$, where $\mathcal{A}$ is the set of characters, $\mathcal{S}$ is a set of semantic markers and $\langle \text{eot} \rangle$ is a special "end of transcription" token. The use of markers from $\mathcal{S}$ allows DAN to go beyond plain HTR, by segmenting the text output into layout elements or named-entities. It even allows increased flexibility: by adapting the ground-truth annotations with custom semantic markers and not necessarily including full transcriptions, one can train DAN to directly address other downstream natural language understanding tasks than those showcased by the authors, such as named entity recognition or key information extraction.

The model we use for fine-tuning has been trained in two phases. First, the network of the encoder and the prediction layer of the decoder have their weights initialized by transfer learning: a simple custom model is built using these modules for the task of text line transcription, and trained on a synthetic dataset of printed lines with the CTC loss. In the second phase, DAN is trained with the cross-entropy loss and teacher forcing on a combination of a real handwritten dataset and of synthetically generated printed documents, with the proportion of synthetic data gradually decreasing. The original authors ran the training on several datasets separately. In each case, synthetic documents are generated with the same layout classes as the dataset, with consistent positioning of the layout elements.

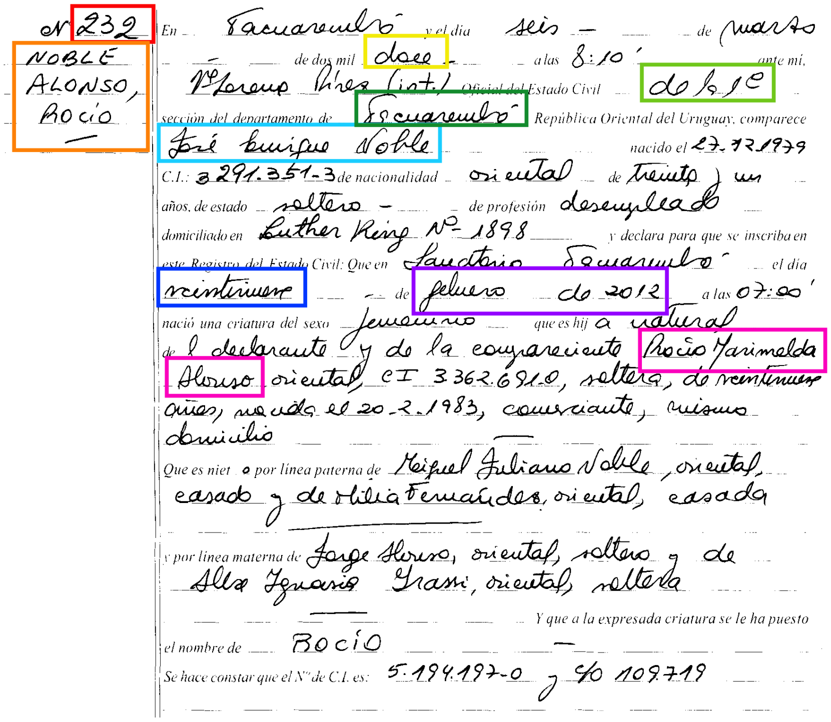
In the present work, we use the model trained in [11] on the RIMES 2009 [14] dataset. This dataset comprises grayscale images of French handwritten letters, with ground-truth annotation consisting in text regions with their layout class (one of "sender", "recipient", "date and location", "subject", "opening", "body" and "PS and attachment") and their text transcription. In order to train DAN, these annotations were automatically translated into sequences of tokens from $\mathcal{D}_{\text{RIMES}} = \mathcal{A}_{\text{RIMES}} \cup \mathcal{S}_{\text{RIMES}} \cup \{\langle\text{eot}\rangle\}$, with consistent rules for reading order determination. For our experiments, we use the open-source implementation of DAN [1].

For our fine-tuning of this model, we need to use a token dictionary $\mathcal{D}_{\text{FT}} = \mathcal{A}_{\text{FT}} \cup \mathcal{S}_{\text{FT}} \cup \{\langle\text{eot}\rangle\}$ which is different from $\mathcal{D}_{\text{RIMES}}$. Indeed, our dataset is not in the same language (Spanish instead of French) thus, while the intersection of character sets $\mathcal{A}_{\text{FT}} \cap \mathcal{A}_{\text{RIMES}}$ is significant, both $\mathcal{A}_{\text{FT}} \setminus \mathcal{A}_{\text{RIMES}}$ and $\mathcal{A}_{\text{RIMES}} \setminus \mathcal{A}_{\text{FT}}$ are non-empty. Furthermore, our fine-tuning task of information extraction is encoded by a distinct set of markers than those for layout segmentation in RIMES, so $\mathcal{S}_{\text{FT}} \cap \mathcal{S}_{\text{RIMES}} = \emptyset$. As a consequence, we need to adapt the token embedding and probability prediction layers in DAN's decoder. Since their weight matrices have a clear interpretation, with individual tokens corresponding to their columns and rows respectively, this does not pose a significant challenge: we keep the trained weights corresponding to tokens in $\mathcal{D}_{\text{FT}} \cap \mathcal{D}_{\text{RIMES}}$, remove those corresponding to $\mathcal{D}_{\text{RIMES}} \setminus \mathcal{D}_{\text{FT}}$ and create new weights for $\mathcal{D}_{\text{FT}} \setminus \mathcal{D}_{\text{RIMES}}$.

4 Datasets

For this work, we had access to a dataset composed of 201 birth certificates from 4 different years (2008, 2012, 2014 and 2016), handwritten in more than 12 writing styles (as can be seen in Figure 3), with their corresponding transcribed text as kept in the computer system used in the local civil registry of Tacuarembó, Uruguay. Images are binarized and saved in tiff format, with a resolution of 200 dpi. The transcription of 7 fields present in the documents was provided in a CSV file. These fields are: "document number", "enrollee's full name", "year of the enrollment", "jurisdiction", "1st parent's full name", "date of birth", and "2nd parent's full name". The field "department" was added to the dataset as a means of having one more annotation with zero annotation cost given that it is constant in all 201 documents. From this original dataset, we created two different sets of annotations.

The first one was obtained by ordering the CSV files containing the original annotations (as given by the civil registry of Tacuarembó) in the document's reading order, and converting the date, originally provided in the format "YYYY-MM-DD", into "date field 1" and "date field 2". The former contains the day of birth written in letters, while the latter contains the month written in letters followed by the word "de" followed by the year written in numbers (YYYY). (For instance, the date "2014-05-31" is annotated as "treinta y uno" in "date field 1" and "mayo de 2014" in the "date field 2".) This decision was made in order to



Nº 232

NOBLE ALONSO, ROCÍO

En Tacuarembó y el día seis — de marzo — de dos mil doce a las 8:10' ante mí, Mª Lorena Pírez (int.) Oficial del Estado Civil de la 1ª sección del departamento de Tacuarembó República Oriental del Uruguay, comparece José Enrique Noble nacido el 27.12.1979 C.I.: 3.291.351-3 de nacionalidad oriental de treinta y un años, de estado soltero — de profesión desempleado domiciliado en Luther King Nº 1898 y declara para que se inscriba en este Registro del Estado Civil: Que en Sanatorio Tacuarembó el día veintinueve de febrero de 2012 a las 07:00' nació una criatura del sexo femenino que es hija natural del declarante y de la compareciente Rocío Marimelda Alonso oriental, CI 3.362.691-0, soltera, de veintinueve años, nacida el 20.2.1983, comerciante, mismo domicilio

Que es nieta por línea paterna de Miguel Juliano Noble, oriental, casado y de Olivia Fernández, oriental, casada

y por línea materna de Jorge Alonso, oriental, soltero y de Alba Ignacia Grassi, oriental, soltera

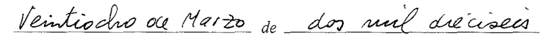
Y que a la expresada criatura se le ha puesto el nombre de ROCÍO

Se hace constar que el Nº de C.I. es: 5.194.197-0 y s/o 109719

Field	Normalized DAN	Diplomatic DAN
document number	232	232
enrollee's full name	Noble Alonso Rocio	NOBLE ALONSO ROCÍO
year of enrollment	doce	doce
jurisdiction	Primera	de la 1a
department	Tacuarembó	Tacuarembó
1st parent's name	José E. Noble	José Enrique Noble
date field 1	veintinueve	veintinueve
date field 2	febrero de 2012	febrero de 2012
2nd parent's name	Rocio M. Alonso	Rocío Marimelda Alonso

Fig. 2: Crop of the upper part of a birth certificate created by copying fields from several real birth certificates originally handwritten by the same public servant (and belonging to the dataset used in experiments), followed by a table showing two mentioned ways of annotating data. Annotated fields are manually highlighted in colors for visualization purposes only: document number (red), enrollee's full name (orange), year of enrollment (yellow), jurisdiction (light green), department (dark green), 1st parent's name (light blue), date field 1 (dark blue), date field 2 (violet), 2nd parent's name (pink).

avoid having to manually annotate the date in each of the documents. However, note that not all handwritten dates are written this way, as Table 1 shows.



Veintiocho de Marzo de dos mil dieciseis

	date field 1	date field 2
Normalized annotation	veintiocho	marzo de 2016
Normalized DAN inference	veintiocho	marzo de 2016
Diplomatic annotation	Veintiocho de Marzo	dos mil dieciseis
Diplomatic DAN inference	Veintiocho de Marzo	dos mil dieciseis

Table 1: Birth certificate crop of a line where the date is handwritten in a way different from the most frequent one, together with annotations for "date field 1" and "date field 2" for both datasets. The inferred text by both models (Diplomatic DAN and Normalized DAN) is also included.

In the provided CSV files, names are also transcribed in (possibly inconsistent) shortened ways. For example, "Alberto Carlos Bustos" can be transcribed as "Alberto C. Bustos" or as "A. Carlos Bustos". As a consequence, data kept in databases are not necessarily verbatim transcriptions of the handwritten text present in the documents. Although it could be ideal if this additional formatting was learned by the model to generate a formatted output (if desired), we decided to create another set of annotations, based on a verbatim transcription of the text (also known as "diplomatic annotation").

This implied creating a new annotation set by: 1) transcribing all accents, when present; 2) respecting upper and lower case; 3) writing each of the two entities corresponding to the birth date in the way they appear in each document (dates can appear as "veintiocho de marzo" and "dos mil dieciseis" in some documents, as in the example shown in Table 1, or as "veintiocho" and "marzo de 2016", which is the most usual way and the one used in the normalized annotation); 4) transcribing names and last names in full, as they appear in the documents (e.g. not abbreviating "Maria Juana González" as "Maria J. González"). Generating the diplomatic transcription of all fields for the 201 documents took about 10 hours of manual work.

Having these two sets of annotations (normalized and diplomatic) enabled us to perform two different experiments, which are described in the next section.

5 Experiments

The main objective of the experiments was to understand the capabilities of DAN for transfer learning, in particular how good this process works in the context of structured birth certificates written in Spanish. We randomly divided the original dataset into train, validation and test in 80% (161 images), 10% (20 images) and 10% (20 images), respectively, and used this same partition for all experiments.

Fine-tuning was done using images of a fixed height of 1900 pixels, automatically resized in such a way that they preserved their aspect ratio. Original images were no smaller than 1800 pixels wide (therefore, they were, at the most, shrunk but never enlarged). The only pre-processing that images went through, before being used in the fine-tuning processes, was a subtle rotation in such a way that the side margin present in all images becomes vertical.

The model features an encoder with 5 convolutional layers and a decoder with 8 transformer layers, each of size 256 with 4 attention heads, employing dropout, data augmentation, and label noise for enhanced performance. The maximum number of epochs is set to 2000, with a batch size of 5 and a learning rate of 0.0001.

We shall now describe our two experiments. The first one consisted in fine-tuning DAN using annotations that required very little pre-processing. We name the generated model **Normalized DAN**. The second one consisted in fine-tuning the same model using diplomatic (verbatim, character by character) annotations. We name the generated model **Diplomatic DAN**.

The reason why we started by training a first model using the annotation dataset obtained directly from the original CSV files (computer-based annotations) is that the use of this kind of annotations is convenient for two reasons. First of all, there is no need for manual transcriptions, which is a costly and time-consuming process. Secondly, the desired result was that the model would output information in a normalized format that can be directly used by a computer system, without post-processing (as it is already formatted with the expected output).

It is worth noting that in this first experiment the model must not only learn how to identify and extract the correct information but also how to format it to the desired output which is, a priori, a more complex task.

The second experiment was designed in order to test how well DAN would perform if fine-tuned with data that reflected exactly what is handwritten in the documents for the same set of fields (i.e. by using a diplomatic annotation).

6 Results

Table 2 shows the Character Error Rate (CER) and the Word Error Rate (WER) for each field and each experiment. We have included two different values for each error measure: regular and normalized text. The first one is considering an exact, case and accent-sensitive comparison between text. For the normalized metric, we have considered uppercase and lowercase to be the same, and also accents were removed. We have included this analysis because, in several occasions, the differentiation of the casing or the presence of accents is not relevant.

A first observation of the quantitative results shown in Table 2 is that both CER and WER values are similar to those obtained in [11], even though the context is different. Comparing the results of the two different models shows that

Field	**Normalized DAN**				**Diplomatic DAN**			
	CER (%)		WER (%)		CER (%)		WER (%)	
	reg.	norm.	reg.	norm.	reg.	norm.	reg.	norm.
document number	3.51	-	10.00	-	3.51	-	10.00	-
year of enrollment	0.00	**0.00**	0.00	**0.00**	3.31	3.31	5.00	5.00
jurisdiction	0.00	**0.00**	0.00	**0.00**	0.92	0.92	3.33	3.33
department	0.00	**0.00**	0.00	**0.00**	4.50	0.00	5.00	0.00
date field 1	0.00	**0.00**	0.00	**0.00**	0.61	0.61	3.85	3.85
date field 2	0.00	**0.00**	0.00	**0.00**	0.35	0.00	1.67	0.00
enrollee’s full name	2.25	2.25	9.64	9.64	2.94	**1.73**	14.46	**6.02**
1st parent’s name	10.72	10.43	36.36	36.36	2.83	**2.02**	12.86	**10.00**
2nd parent’s name	15.81	15.81	35.19	35.19	7.17	**5.70**	28.36	**20.90**
global	4.36	4.32	13.92	13.92	3.05	**2.04**	11.27	**7.51**

Table 2: Character Error Rate and Word Error Rate for each field of birth certificates in the test set (composed of 20 birth certificate scans). Both metrics are computed on regular text (reg.) and normalized text (norm.), in which characters are changed to lowercase and accents are removed.

Diplomatic DAN globally performs slightly better than **Normalized DAN**. Yet, a closer look at the results obtained for each field reveals that **Normalized DAN** performs better than **Diplomatic DAN** in most fields (“year of enrollment”, “jurisdiction”, “department”, “date field 1” and “date field 2”), while **Diplomatic DAN** performs considerably better in fields that contain names and last names (“enrolee’s full name”, “1st parent’s name” and “2nd parent’s name”). We analyze the outputs based on this division of fields.

Fields that do not contain names and last names Six out of the nine fields do not contain names or last names. In these cases, the results suggest that giving the model a normalized annotation to learn (which is what was used for Normalized DAN), improves the model’s learning capabilities. The number of errors that Diplomatic DAN makes in all normalizable fields is shown in Table 3. Even if the results obtained by the two models can be comparable, training the **Normalized model** is less costly in terms of annotation effort and it also generates a formatted output, which is usually preferred when saving data into information systems.

Fields that contain names and last names As mentioned before, there are three fields in the dataset that contain names and last names: “enrollee’s full name”, “1st parent’s name” and “2nd parent’s name”. In the three cases, the CER and WER, if compared ignoring cases and accents, are better with **Diplomatic DAN**. It is interesting to see how this improvement is different in the tree mentioned fields.

The field “enrollee’s full name” is only slightly better for Diplomatic DAN than for Normalized DAN. It is worth remembering that, in this specific field,

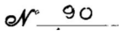
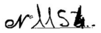
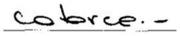
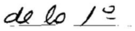
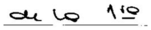
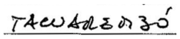
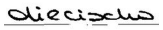
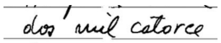
Field	Annotation	Inference	Crops from documents
document number	**9**0	**2**0	 Nº 90
	115**4**	115**6**	 nº1154
year of enrollment	catorce	c**o**ce**r**	 catorce.-
jurisdiction	de la 1**o**	de la 1**a**	 de la 1º
	de la 1r**a**	de la 1r**_**	 de la 1º
department	T**ACUAREMBÓ**	T**acuarembó**	 TACUAREMBÓ
date field 1	dieci**o**cho	dieci**s**cho	 dieciocho
date field 2	**c**atorce	**C**atorce	 dos mil catorce

Table 3: Errors made by Diplomatic DAN on normalizable fields. The only errors made by Normalized DAN in all normalizable fields are also shown in the table (they are the ones made in document number). Errors made when an uppercase letter instead of its corresponding lowercase letter (or vice-versa) are marked in blue. Errors made when actually substituting one character for another one or omitting a character are marked in red. No insertion errors were detected in these fields.

where names and last names typically appear in several lines, in the margin of the certificate, a full annotation is provided both in the normalized dataset and in the diplomatic one. This probably explains the very similar results that are obtained by both models.

The main difference between the "1st parent's name" and "2nd parent's name" fields is that the former is surrounded by preprinted text while the latter is always surrounded by handwritten text. This probably explains why the behavior of Diplomatic DAN in the "2nd parent's name" field is almost 3 times better than Normalized DAN, while the improvement obtained by Diplomatic DAN on the "1st parent's field" is more than 5 times better (when evaluated ignoring cases and accents).

One of the possible reasons why Diplomatic DAN works better than Normalized DAN on these three fields was already explained in the description of the dataset: annotations provided by the civil registry included some abbreviation criteria on both parents' names fields where, very frequently, the second name was abbreviated to its initial (e.g., "Alberto Carlos Bustos" was transcribed as "Alberto C. Bustos"). Somehow, this was learned by the model but with mixed results. In cases where the enrollee has a second name and the model was able to understand which letter to use, Normalized DAN produced a correct output (i.e. correctly abbreviated the middle name). In other cases, it made a mistake when abbreviating the middle name. In cases where no middle name was handwritten in the document, the model just invented one, presumably to agree with the format of the data it was trained on.

7 Conclusions and future work

In this work we explored the capabilities of fine-tuning DAN using two annotation strategies to adapt it to a different context than the one in which it was trained. We observed that, even with very little data (181 documents), the model was able to obtain very similar results to the ones shown in the original paper.

We also explored two annotation strategies, normalized and diplomatic, and saw that the former worked better in fields that can be standardized while the latter performed better in fields where character by character transcription is needed. The aforementioned results on fine-tuning DAN for transcribing handwritten documents open several questions.

The first one is whether using a "hybrid" annotated dataset (meaning, normalized for normalizable fields and diplomatic for fields containing names and last names) would result in a model that is capable of both transcribing "normalizable" fields in a normalized way (hence, eliminating the need of post-transcription normalization) while being extremely accurate when transcribing non-normalizable fields (such as, in our case, names and last names).

A second natural line is exploring how far we can go in reducing the number of birth certificates needed to fine-tune a model without losing accuracy.

A third line we will explore is the model's generalization capabilities. For instance, when dealing with different document layouts. Related to this, it is worth testing whether using a different annotation order for birth certificates that have the margin on the left or on the right will improve accuracy.

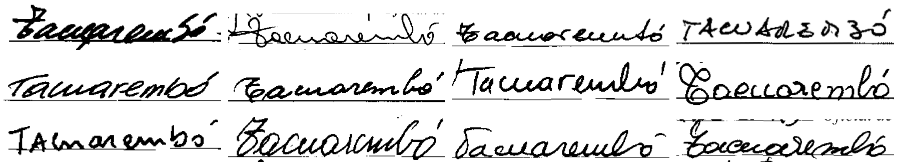


Fig. 3: Images of the word "Tacuarembó" as written by 12 different writers.

Further, one can explore whether exploiting document redundancy by annotating redundant information will result in higher accuracy. As can be seen from Figure 2, birth certificates generally have some data redundancy such as the enrolled person's name is written both in the margin and in the body of the document. Lastly, since handwriting styles can be so different among one another, we would like to explore whether training a model on a specific writer can lead to a better understanding of the writer's specific handwriting style and, therefore, a higher transcription accuracy.

Acknowledgments. The research that originated the results presented in this publication was partly supported by the Agencia Nacional de Investigación e Innovación (ANII) and the France 2030 CollabNext project.

References

1. Dan implementation repository by teklia. https://gitlab.teklia.com/atr/dan, release: 0.2.0rc6
2. Abadie, N., Carlinet, E., Chazalon, J., Duménieu, B.: A benchmark of named entity recognition approaches in historical documents application to 19 th century french directories. In: International Workshop on Document Analysis Systems. pp. 445–460. Springer (2022)
3. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: Flair: An easy-to-use framework for state-of-the-art nlp. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations). pp. 54–59 (2019)
4. Arora, A., Chang, C.C., Rekabdar, B., BabaAli, B., Povey, D., Etter, D., Raj, D., Hadian, H., Trmal, J., Garcia, P., et al.: Using asr methods for ocr. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 663–668. IEEE (2019)
5. Bluche, T., Louradour, J., Messina, R.: Scan, attend and read: End-to-end handwritten paragraph recognition with mdlstm attention. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1, pp. 1050–1055. IEEE (2017)
6. Boillet, M., Tarride, S., Schneider, Y., Abadie, B., Kesztenbaum, L., Kermorvant, C.: The socface project: Large-scale collection, processing, and analysis of a century of french censuses (2024)
7. Cheplygina, V., Varoquaux, G.: Artificial intelligence in science: Lessons from shortcomings in machine learning for medical imaging. In: Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research. Organization for Economic Co-operation and Development (OECD) (2023)
8. Clérice, T., Ariane, P., Malamatenia, V.E., Alix, C., Jean-Baptiste, C., Matthias, G.L., Olivier, B.F., Franz, F., Michaels, G., Agnès, B., et al.: Catmus medieval: A multilingual large-scale cross-century dataset in latin script for handwritten text recognition and beyond (2024)
9. Constum, T., Kempf, N., Paquet, T., Tranouez, P., Chatelain, C., Brée, S., Merveille, F.: Recognition and information extraction in historical handwritten tables: Toward understanding early 20th century paris census. In: Uchida, S., Barney, E., Eglin, V. (eds.) Document Analysis Systems. pp. 143–157. Springer International Publishing, Cham (2022)
10. Coquenet, D., Chatelain, C., Paquet, T.: End-to-end handwritten paragraph text recognition using a vertical attention network. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 508–524 (2022)
11. Coquenet, D., Chatelain, C., Paquet, T.: Dan: a segmentation-free document attention network for handwritten document recognition. IEEE transactions on pattern analysis and machine intelligence (2023)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

13. Graves, A., Schmidhuber, J.: Offline handwriting recognition with multidimensional recurrent neural networks. In: Proceedings of the 21st International Conference on Neural Information Processing Systems. p. 545–552. NIPS'08, Curran Associates Inc., Red Hook, NY, USA (2008)
14. Grosicki, E., Carré, M., Brodin, J.M., Geoffrois, E.: Results of the rimes evaluation campaign for handwritten mail processing. In: 2009 10th International Conference on Document Analysis and Recognition. pp. 941–945. IEEE (2009)
15. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia. p. 4083–4091. MM '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3548112, https://doi.org/10.1145/3503161.3548112
16. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 498–517. Springer Nature Switzerland, Cham (2022)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
18. Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: Trocr: Transformer-based optical character recognition with pre-trained models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 13094–13102 (2023)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems **26** (2013)
20. Monnier, T., Aubry, M.: docextractor: An off-the-shelf historical document element extraction. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 91–96. IEEE (2020)
21. Monroc, C.B., Miret, B., Bonhomme, M.L., Kermorvant, C.: A comprehensive study of open-source libraries for named entity recognition on handwritten historical documents. In: International Workshop on Document Analysis Systems. pp. 429–444. Springer (2022)
22. Nion, T., Menasri, F., Louradour, J., Sibade, C., Retornaz, T., Métaireau, P.Y., Kermorvant, C.: Handwritten information extraction from historical census documents. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 822–826. IEEE (2013)
23. Oliveira, S.A., Seguin, B., Kaplan, F.: dhsegment: A generic deep-learning approach for document segmentation. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 7–12. IEEE (2018)
24. Peng, Q., Pan, Y., Wang, W., Luo, B., Zhang, Z., Huang, Z., Cao, Y., Yin, W., Chen, Y., Zhang, Y., Feng, S., Sun, Y., Tian, H., Wu, H., Wang, H.: ERNIE-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 3744–3756. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.findings-emnlp.274, https://aclanthology.org/2022.findings-emnlp.274

25. Petitpierre, R., Kramer, M., Rappo, L.: An end-to-end pipeline for historical censuses processing. International Journal on Document Analysis and Recognition (IJDAR) **26**(4), 419–432 (2023)
26. Puigcerver, J.: Are multidimensional recurrent layers really necessary for handwritten text recognition? In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1, pp. 67–72. IEEE (2017)
27. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. arXiv preprint arXiv:2003.07082 (2020)
28. Romero, V., Fornés, A., Serrano, N., Sánchez, J.A., Toselli, A.H., Frinken, V., Vidal, E., Lladós, J.: The esposalles database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognition **46**(6), 1658–1669 (2013). https://doi.org/https://doi.org/10.1016/j.patcog.2012.11.024, https://www.sciencedirect.com/science/article/pii/S0031320312005080
29. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
30. Singh, S.S., Karayev, S.: Full page handwriting recognition via image to sequence extraction. In: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part III 16. pp. 55–69. Springer (2021)
31. Tarride, S., Boillet, M., Kermorvant, C.: Key-value information extraction from full handwritten pages. In: International Conference on Document Analysis and Recognition. pp. 185–204. Springer (2023)
32. Tarride, S., Boillet, M., Moufflet, J.F., Kermorvant, C.: Simara: A database for key-value information extraction from full-page handwritten documents. In: International Conference on Document Analysis and Recognition. pp. 421–437. Springer (2023)
33. Tarride, S., Lemaitre, A., Coüasnon, B., Tardivel, S.: A comparative study of information extraction strategies using an attention-based neural network. In: International Workshop on Document Analysis Systems. pp. 644–658. Springer (2022)
34. Tarride, S., Maarand, M., Boillet, M., McGrath, J., Capel, E., Vézina, H., Kermorvant, C.: Large-scale genealogical information extraction from handwritten quebec parish records. International Journal on Document Analysis and Recognition (IJDAR) **26**(3), 255–272 (Sep 2023). https://doi.org/10.1007/s10032-023-00427-w
35. Tu, Y., Guo, Y., Chen, H., Tang, J.: Layoutmask: Enhance text-layout interaction in multi-modal pre-training for document understanding. In: Annual Meeting of the Association for Computational Linguistics (2023), https://api.semanticscholar.org/CorpusID:258967524
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
37. Wigington, C., Tensmeyer, C., Davis, B., Barrett, W., Price, B., Cohen, S.: Start, follow, read: End-to-end full-page handwriting recognition. In: Proceedings of the European conference on computer vision (ECCV). pp. 367–383 (2018)