Information Extraction in Handwritten Marriage Licenses Books

Verónica Romero[1], Alicia Fornés[2], Emilio Granell[1], Enrique Vidal[1], Joan Andreu Sánchez[1]

[1] PRHLT Research Center, Universitat Politècnica de València, Spain

{vromero,egranell,evidal,jandreu}@prhlt.upv.es

[2] Computer Vision Center, Dept. Computer Science, Universitat Autónoma de Barcelona, Spain

afornes@cvc.uab.es

ABSTRACT

Handwritten marriage licenses books are characterized by a simple structure of the text in the records with an evolutionary vocabulary, mainly composed of proper names that change along the time. This distinct vocabulary makes automatic transcription and semantic information extraction difficult tasks. Previous works have shown that the use of category-based language models and a Grammatical Inference technique known as MGGI can improve the accuracy of these tasks. However, the application of the MGGI algorithm requires an a priori knowledge to label the words of the training strings, that is not always easy to obtain. In this paper we study how to automatically obtain the information required by the MGGI algorithm using a technique based on Confusion Networks. Using the resulting language model, full handwritten text recognition and information extraction experiments have been carried out with results supporting the proposed approach.

CCS CONCEPTS

• **Information systems** → Information retrieval; • **Computing methodologies** → Machine learning.

KEYWORDS

Handwritten recognition, Marriage Licenses, Information extraction

ACM Reference Format:

Verónica Romero[1], Alicia Fornés[2], Emilio Granell[1], Enrique Vidal[1], Joan Andreu Sánchez[1]. 2018. Information Extraction in Handwritten Marriage Licenses Books. In *HIP 19: International Workshop on Historical Document Imaging and Processing, September 20–21, 2019, Sydney, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Handwritten marriage licenses books [13] have been used for centuries by ecclesiastical and secular institutions to register marriages. The information contained in these historical documents, together with other population documents such as birth, death and census records [9], provide insights of the life of our ancestors. Indeed, this type of population documents contain relevant semantic information that is useful for scholars for studying demographic and social changes, migration movements, genealogical research, etc. Therefore, one of the goals of this kind of documents, rather than to transcribe them perfectly, is to extract their relevant information to allow scholars and citizens to make use of this information through semantic searches.

It must be noted that, if a perfect transcription can be obtained, then identifying the relevant semantic information (e.g. named entities) is much easier. For example, Natural Language Processing (NLP) techniques could be applied after obtaining the transcribed text. However, given the difficulties inherent in historical manuscripts, their transcription is not accurate. Luckily, it is not mandatory to obtain a perfect transcription of all the manuscript, because only the relevant words (e.g. family names, places, dates, etc.) are required for filling the knowledge databases, which will be later accessed by either scholars or society in general.

The automatic transcription of historical documents is typically based on techniques that have been used in Automatic Speech Recognition (ASR), such as Hidden Markov Models (HMM) [3, 18], (Multi-dimensional) Recurrent Neural Networks (ANN) [6, 12] and attention models [1].

As in speech recognition, the language model plays a fundamental role in Handwritten Text Recognition (HTR) by adequately restricting the search space and solving ambiguities in the recognition. Indeed, the above mentioned HTR methods often incorporate dictionaries and language models based on n-grams or recurrent neural networks to improve the performance.

In the case of marriage records, the state-of-the-art approaches typically use two separate subsequent tasks: first transcription and second named entities recognition, as it is shown at the IEHHR competition [4]. Only recently, some approaches for a joint transcription and named entity recognition have been proposed [2, 11].

However, and given the regular structure of marriage records, we claim that specific language models can help to better extract not only the transcription, but also the semantic information. For example, in [15], a category-based language model [10] was successfully used both to better represent the regularities in marriage license books and to obtain the relevant words in each record. In [17], a language model based on recurrent neural networks was proposed to model the sequence of categories and improve the extraction of named entities. In [14], a Morphic Generator Grammatical Inference (MGGI) technique [5, 19] was proposed to improve the semantic accuracy of category-based language models. In MGGI, a priori knowledge is used to label (some of) the words of the training strings and then, a bi-gram can be trained on these transformed strings. In

HIP 19, September 20–21, 2019, Sydney, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00

https://doi.org/10.1145/1122445.1122456

this way, the syntactic dependencies, which according to the prior knowledge are considered important to understand the handwritten record, become reflected in the language model produced by MGGI.

However, the required a priori knowledge may not be available or difficult to obtain. Generally, it is necessary to analyze the errors occurred using a conventional language model, and then, relabel the words of the training strings in order to convey language constraints that help avoiding the observed errors. Such an analysis requires familiarity on the task and also on the language of the considered document, which may lead to difficulties in the application of the MGGI technology.

In this paper, rather than focusing on the optical model used for the transcription, we focus on improving the language model for semantic information extraction. To overcome the above mentioned limitation, we propose a technique to discover the dependencies required to relabel the training strings in an automatic way. We use a method based on the construction of confusion networks. From the resulting confusion network, words with higher probability are used as anchors to automatically relabel the training strings. Thanks to this technique, human intervention is no longer needed to provide the required prior knowledge. As a result, the MGGI can be applied regardless of the language or the content of the document.

The rest of the paper is organized as follows. Section 2 describes the handwritten marriage license book collection and the information extraction task. Then, the category based language model and the MGGI techniques are introduced in Sections 3 and 4, respectively. The method based on confusion networks to discover the prior knowledge is explained in 5. The experimental framework and results are presented in Sections 6 and 7. Finally, the conclusions and the future work are explained in the last section.

2 TASK DESCRIPTION

A handwritten marriage license book from a collection conserved at the Archives of the Cathedral of Barcelona [13] has been used. A page example of this manuscript is shown in Figure 1.

Each page is divided horizontally into three blocks, the husband surname's block (left), the main block (middle) and the fee block (right). Vertically, the page is divided into individual license records. Each marriage license typically contains information about the marriage day and groom's and bride's information. This information is written following an order: the groom's information is written first and then the bride's information. Inside the groom's part and the bride's part, the information is written following an order: the given name and surnames, occupation, parents' information, etc. In some cases, additional information is given or some information is missing. Thus, the vocabulary changes along the license: the first part is related to the groom, with names related to men and occupations, whereas, the last part is the bride's part. Fig. 2 shows an example of a marriage license whose transcription is:

```
Dit dia rebere$ de Raphel Joani texidor de lli de
Vilassar fill de Miquel Joani texidor de lli y de
Violant, ab Sperensa do$sella filla de Sebastia Garau
Pere Boter de dita parrochia y de t.
```

It must be noted that a common problem when transcribing handwritten marriage license books by means of HTR methods is that the

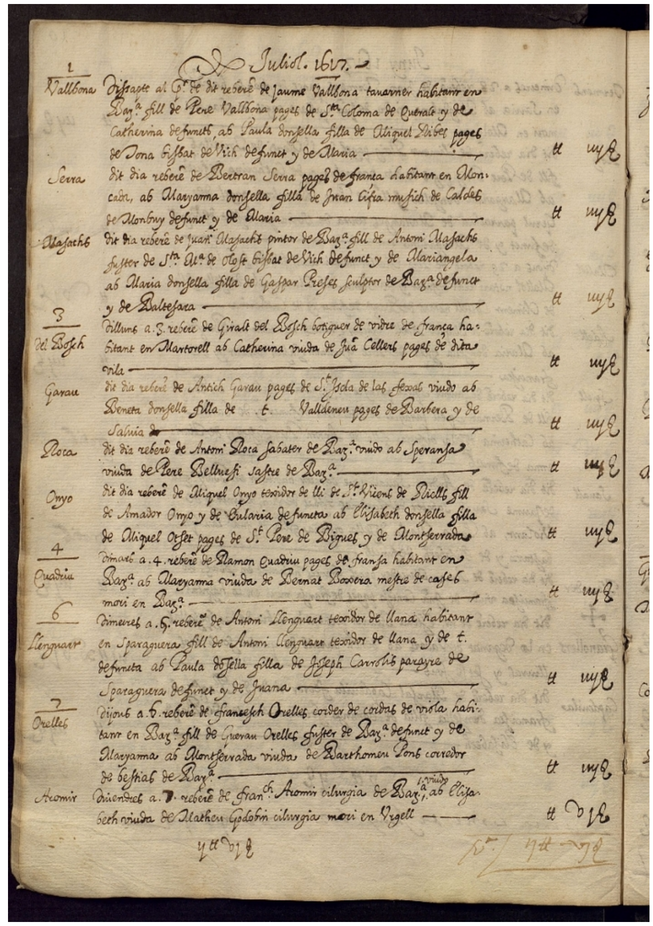

Figure 1: Example of a marriage license page. Each page is divided horizontally into three blocks, the husband surname's block (left), the main block (middle) and the fee block (right)

classical n-gram language models can be very inaccurate due to the restrictions of the language underlying and also due to the special vocabulary of the task, which is composed mainly of proper nouns. In [15, 17] it has been shown that the use of category based language models can generalize the word patterns that have never been seen during training with good results.

3 CATEGORY-BASED HTR

As discussed before, the use of a category language model in HTR can benefit both, the handwritten accuracy and the semantic information extraction process.

This improvement is due to two main reasons. Firstly, given that category-based language models share statistics between words of the same category, category-based models are able to generalize to word patterns that are never encountered in the training corpus. Secondly, grouping words into categories can reduce the number of contexts in an n-gram model, and thereby reduce the training set sparseness problem.

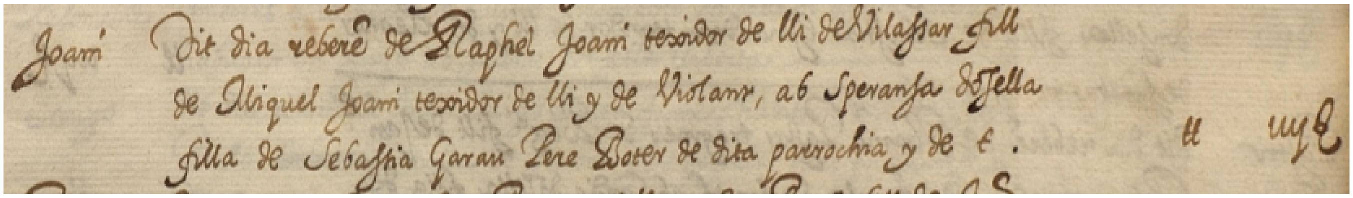

Figure 2: Example of a marriage license.

In this work, the same semantic categories defined in [15] have been used: groom's (*Gr*) given name and surname, bride's (*Br*) given name and surname, parents' (*Fa* and *Mo*) given names and surnames, occupations (*Oc*), place of residence (*Resi*), geographical origin, etc. Then, a category-based language model has been generated and integrated into the handwritten text recognition process.

The proposed categorization involves classifying only some words in the vocabulary and not all of them. In this way, a partially categorized corpus was obtained. Words without a category could be viewed as categories that contain a single word.

Formally speaking, let $\mathbf{x}$ be a handwritten sentence image, let $\mathbf{w}$ be a word sequence, and let $\mathbf{c}$ be the sequence of categories associated to the word sequence. Following the discussion presented in [14], the Viterbi decoding process can yield not only the best word sequence hypothesis, $\hat{\mathbf{w}}$, but also the corresponding best sequence of semantic categories, $\hat{\mathbf{c}}$:

$$(\hat{\mathbf{c}}, \hat{\mathbf{w}}) \approx \arg\max_{\mathbf{c},\mathbf{w}} p(\mathbf{x} \mid \mathbf{w}) \cdot p(\mathbf{w} \mid \mathbf{c}) \cdot p(\mathbf{c}) \tag{1}$$

$P(\mathbf{x} \mid \mathbf{w})$ represents the optical-lexical knowledge and is typically approximated by concatenating character models such as HMMs [7]. $P(\mathbf{w} \mid \mathbf{c})$ is the word-category distribution and is usually approximated by an 1-gram for each category. Finally, $p(\mathbf{c})$ is the probability of a category sequence and is approximated by an n-gram.

4 LANGUAGE MODELING USING MGGI

It is well known that n-gram models are just a subclass of probabilistic finite-state machines (PFSM) [20]. Therefore the capabilities of n-grams to model relevant language contexts or restrictions is limited, not only with respect to more powerful syntactic models such as context-free grammars, but also even with respect to the general class of PFSMs. In fact, no n-gram can properly approach (word) string distributions involving long-span dependencies which are common in natural language [20]. For instance, no n-gram (with bounded n) can approach a distribution of strings over the vocabulary $\{a, b, c, d, e\}$ such that the probability is high for the strings ab^ic or db^ie and is low or null for other strings such as ab^ie and db^ic, where i is any arbitrarily large integer. However, such a distribution can be exactly modeled by a very simple PFSM (see [20], Sec. 2.1.3).

While learning PFSMs from training strings is in general hard, there is a framework which allows to learn PFSMs which can model *given*, albeit arbitrarily complex finite-state restrictions. This framework, known as "Morphic Generator Grammatical Inference" (MGGI), provides a methodology for using prior knowledge about important task constraints, to ensure that the trained finite-state models will comply with these constraints.

In MGGI, the a priori knowledge is used to label (some of) the words of the training strings and a simple bigram is trained on the transformed strings. Then an inverse transformation is applied to this bigram to obtain a PFSM which deals with the restrictions conveyed by the initial string transformation [5, 20]. A direct applications of these ideas to build accurate PFSM language models for automatic speech recognition can be seen in [19]. In [14], the MGGI was applied to the recognition task of a handwritten marriage license book. In that work, the labelling used in the MGGI intend to solve the mis-categorization of the bride's family information as groom's information, due to a wrong bigram generalization.

However, this a priori information was manually provided. In the next subsection, we describe the automatic discovery of this a priori knowledge through Confusion Networks.

5 AUTOMATIC PRIOR KNOWLEDGE EXTRACTION

As previously commented, the MGGI approach requires a priori knowledge to label the training strings. In the task we are considering, this knowledge consists of the words that allow to separate the marriage licenses in the different categories used in the MGGI approach.

In [14], this prior knowledge was manually obtained by analyzing the errors produced using standard n-gram category-based language models. It required to find "anchor words" to separate a marriage license into two parts: the groom's information and the bride's information. The aim was to solve the mis-categorization of the bride's family information as groom's information, due to wrong n-gram generalizations.

To overcome this limitation, we get inspiration from the work presented in [8] for finding consensus using confusion networks. Thus, we follow this approach and study how to automatically obtain the anchor words required for the MGGI labeling by using a confusion network obtained from the training samples.

A Confusion Network (CN) is a compact topology used to represent lattices. It is a weighted directed graph, in which each path goes through all the nodes. The words and their probabilities are stored in the edges, and the total probability of the words contained in all edges between two consecutive nodes sum up to 1.

To build a CN from a set of strings, the strings are aligned by clustering the words according to their similarity and occurrence time. Then, for each cluster and for each different word, the word probability is obtained as the counts of this word in the cluster over the total number of words in the cluster. Finally, the CN is composed of these clusters in order. These operations can be easily performed using the SRILM toolkit [16]. Figure 3 shows an example of CN obtained from the following set of marriage license (sub)strings:

```
- dit dia rebere$ de Jaume Salavert
parayre habitant en Terrassa ...
- rebere$ de Onofre Tapies # de Terrassa
- dit dia rebere$ de Pere Cugul
treballador habitant en Bar^(a). ...
- rebere$ de Miquel Tintorer parayre de
Bar^(a). ...
- dit dia rebere$ de francesch Nogues
corder viudo de Mataro ...
```

As it can be observed in Figure 3, the words that are repeated in the vast majority of licenses, such as "rebere$" have the highest probabilities.

In this work, we combine all the training sentences and then look for the words with highest probability to be used as anchors in the new labelling. Without taking into account the stop words, the most probable word in the resulting CN is *ab*. Note that this word does not appear in Figure 3 because in that figure only partial marriage licenses, where the word *ab* is not included, are used. Not surprisingly, this is the same word that has been used as anchor in [14]. This means that by using our automatic method, and choosing only the most probable word, the same results as in [14] can be obtained.

In this work we can improve these results by choosing the three most frequent words, which corresponds to: *ab*, *rebere$* and *filla*. Therefore, we relabel the training samples using these words.

In the vast majority of the records, the groom's information is introduced by the word *rebere$*, the groom's and the bride's information is separated by the word *ab*, and the word *filla* is used to separate the bride's and her parents information. Using these words, it is straightforward to label all the tokens between the word *"rebere$"* and the word *"ab"* with the suffix *"G"*, those appearing between *"ab"* and *"filla"* as *"B"* and those appearing after *"filla"* as *"A"*. By applying this labeling scheme to the categorized training transcripts of the license of the Figure 2, the following training text is obtained:

```
Dit dia rebere$G deG [GrName]G
[GrSurname]G [GrProf]G deG [GrResi]G
fillG deG [GrFaName]G [GrFaSurname]G
[GrFaProf]G yG deG [GrMoName]G
,G ab [BrName]B do$sellaB fillaA deA
[BrFaName]A [BrFaSurname]A [BrFaProf]A
deA [BrFaResi]A yA deA [BrMoName]A
```

After training a category-based bigram, the inverse transformation required by MGGI just consists in removing these suffixes. The resulting PFSM adequately models the dependencies conveyed by the labeling adopted.

6 EXPERIMENTAL FRAMEWORK

In this section we describe the corpus, the system setup and the metrics that are used to evaluate our approach.

6.1 Corpus

In the experiments, we have used the ESPOSALLES database [13], consisting in one marriage license book from the Cathedral of Barcelona. The corpus, written by one single writer in old Catalan in the 17th century, is composed of 173 pages, 5,447 lines grouped in 1,747 licenses. The volume contains around 60,000 running words from a lexicon of around 3,500 different words. An expert paleographer transcribed the manuscript and annotated the semantic information according to the 40 different categories defined by expert demographers, as described in [15].

For the sake of comparison with the work carried out in [14], we use the same seven partitions of 25 pages defined in [13] for cross-validation experiments. Table 1 shows the average values of the statistics related with the different partitions.

Table 1: Basic statistics of the database and average values from the 7 different partitions.

Number of:	Total	Average
Pages	173	24.7
Licenses	1,747	249.6
Lines	5,447	778.1
Run. words	60,777	8682.4
OOV	–	361
Lexicon	3465	1070
Characters	328229	46889.9
Semantic labels	21386	3055.1

6.2 System setup

We carried out seven rounds, with each of the partitions used once as test data and the remaining six partitions used as training data.

The pages were divided into line images, and normalized as explained in [13]. For each line image, we extracted a sequence of feature vectors [18] based on the gray level of the image. Since we carried out experiments at license level, the feature sequences of the lines have been concatenated into licenses. Note that the same procedure has also been followed in some other works, such as in [2].

The characters were modeled by continuous density left-to-right HMMs with 6 states and 64 Gaussian mixture components per state. These parameters worked well in previous handwriting recognition experiments. These models were estimated by training text images represented as feature vector sequences using the Baum-Welch algorithm. For decoding we used the Viterbi algorithm [7]. A category-based bi-gram was estimated using the MGGI methodology from the training transcriptions of the text line images. Those words in the test partition that do not appear in the training partition, named Out of Vocabulary (OOV) words, were added as singletons to the corresponding word category distribution. For OOV words that belong to a category that has not been seen during the training step, we add the category in the category-based 1-gram and the word in the category distribution as singleton. The word category distributions were modeled by uni-grams.

6.3 Assessment Measures

To assess the quality of the transcription, we use the *Word Error Rate* (WER), defined as the minimum number of words that need to be substituted, deleted or inserted to convert the sentences recognized by the system into the reference transcriptions, divided by the total number of words in these transcriptions.

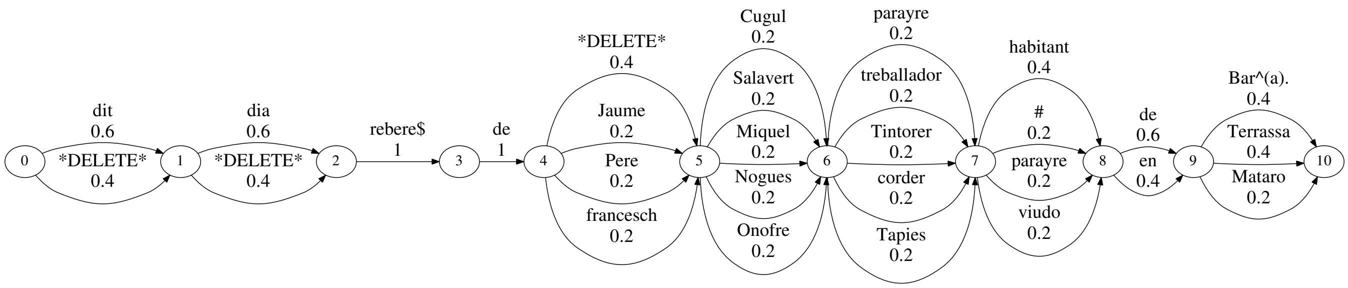

Figure 3: Example of the Confusion Network obtained from a set of strings of partial marriage licenses.

To asses the quality of the information extraction, we use the precision and recall measures, defined in terms of the number of relevant words. Relevant words are the ones that belong to any of the 40 defined categories. For instance, in the example shown in previous sections, the relevant words are those associated to a category: *Raphel, Joani, texidor_de_lli, Vilassar, ...* Formally, let R be the number of relevant words contained in the document, D the number of relevant words that the system has detected, and C the number of relevant words correctly detected by the system. Precision π) and recall (ρ) are computed as:

$$\pi = \frac{C}{D} \; \rho = \frac{C}{R}$$

7 RESULTS

The results are shown in Table 2. Our approach has been compared to a category-based (CB) 2-gram language model (LM) [15] and to the MGGI work presented in [14]. The mean Precision and Recall are computed for the absolute number of instances (I) or by averaging the Precision and Recall for each one of the categories (C). Obviously, the absolute values are higher because the categories are not balanced. Some of them only appear in few licenses, and consequently, the ability of the model to learn is lower.

As stated in section 5, if the MGGI uses only one anchor word, the results obtained with the automatic method presented in this paper are the same that those obtained in [14] (second row of Table 2), given that the detected anchor word is exactly the same (*"ab"*). Instead, if we use three automatically detected anchor words, we can observe an improvement (row MGGI+CN). Although we use cross-validation, it must be noted that the syntactical structure is similar, so the three detected anchor words have been the same in all the folds.

Concerning the transcription accuracy, the WER slightly decreases using the three anchors. From the 60.000 words, now the system could correctly transcribe 36 words more. Note that the same optical models are used in [14] and in this paper and the application of the MGGI intend to solve the mis-categorization problems. This can explain the slight improvement at WER level. In case of the information extraction task, the improvement in Precision and Recall is significant.

It is important to note that we consider an error whenever the semantic category or the transcription are incorrect. If a word transcription is incorrect, it will also be a semantic labeling error, no matter if the category is correct. Consequently, the semantic labeling error is never lower than the WER.

Table 2: WER, precision (π) and recall (ρ) obtained with the Category-Based LM (CB), the MGGI and the MGGI with Confusion Networks (MGGI+CN). The mean is computed for the absolute number of instances (I) and for categories (C). All results are percentages.

	WER	I-π	I-ρ	C-π	C-ρ
CB [15]	10.1	79.2	66.6	73.5	65.2
MGGI [14]	10.1	85.3	76.2	78.3	72.2
Our work: MGGI + CN	10	87.2	77.7	80.3	73.4

8 CONCLUSIONS

In this paper we have argued that specific language models can help in semantic recognition. Moreover, we have shown how Confusion Networks can automatically detect the anchor words that are required for the MGGI methodology. Thanks to the resulting language model, the information contained in structured documents such as marriage licenses, can be automatically transcribed and semantically labeled without the intervention of an expert user. In the near future, we would like to extend our work to other kind of structured documents.

Even though in this paper we have focused on improving the language model, we agree that more sophisticated optical models for text recognition will also help in the overall information extraction performance. Therefore, as a future work we also plan to carry out experiments using Convolutional Recurrent Neural Networks for optical modelling, which have recently shown notable improvements in the HTR accuracy.

9 ACKNOWLEDGMENTS

Work partially supported by the BBVA Foundation through the 2017–2018 Digital Humanities research grant "Carabela", by Ministerio de Ciencia/AEI/FEDER/EU through the coordinated MIRANDA project (RTI2018-095645-B-C21 and RTI2018-095645-B-C22), by EU JPICH project "HOME – History Of Medieval Europe" (Spanish PEICTI Ref. PCI2018-093122) and by the Ramon y Cajal Fellowship RYC-2014-1683

REFERENCES

[1] Théodore Bluche, Jérôome Louradour, and Ronaldo Messina. 2017. 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. IEEE, 1050–1055.

[2] Manuel Carbonell, Mauricio Villegas, Alicia Fornés, and Josep Lladós. 2018. Joint recognition of handwritten text and named entities with a neural end-to-end model. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 399–404.

[3] Andreas Fischer, Volkmar Frinken, and Horst Bunke. 2013. Hidden markov models for off-line cursive handwriting recognition. In *Handbook of Statistics*. Vol. 31. Elsevier, 421–442.

[4] Alicia Fornés, Verónica Romero, Arnau Baró, Juan Ignacio Toledo, Joan Andreu Sánchez, Enrique Vidal, and Josep Lladós. 2017. ICDAR2017 competition on information extraction in historical handwritten records. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1. IEEE, 1389–1394.

[5] P. Garcia, E. Vidal, and F. Casacuberta. 1987. Local languages, the succesor method, and a step towards a general methodology for the inference of regular grammars. *IEEE Transactions on PAMI* 6 (1987), 841–845.

[6] Alex Graves and Jürgen Schmidhuber. 2008. Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks.. In *NIPS*. 545–552.

[7] F. Jelinek. 1998. *Statistical Methods for Speech Recognition*. MIT Press.

[8] L. Mangu, E. Brill, and A. Stolcke. 2000. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language* 14, 4 (2000), 373–400.

[9] Joan Mas, Alicia Fornés, and Josep Lladós. 2016. An interactive transcription system of census records using word-spotting based information transfer. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. IEEE, 54–59.

[10] T.R. Niesler and P.C. Woodland. 1996. A variable-length category-based n-gram language model. In *Proc. of ICASSP-96*, Vol. 1. 164 –167 vol. 1. https://doi.org/10.1109/ICASSP.1996.540316

[11] Animesh Prasad, Hervé Déjean, Jean-Luc Meunier, Max Weidemann, Johannes Michael, and Gundram Leifert. 2018. Bench-marking information extraction in semi-structured historical handwritten records. *arXiv preprint arXiv:1807.06270* (2018).

[12] Joan Puigcerver. 2017. Are multidimensional recurrent layers really necessary for handwritten text recognition?. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1. IEEE, 67–72.

[13] V. Romero, A. Fornés, N. Serrano, J. A. Sánchez, A.H. Toselli, V. Frinken, E. Vidal, and J. Lladós. 2013. The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. *Pattern Recognition* 46 (2013), 1658–1669.

[14] V. Romero, A. Fornés, E. Vidal, and J.A. Sánchez. 2016. Using the MGGI Methodology for Category-based Language Modeling in Handwritten Marriage Licenses Books. In *ICFHR*. Shenzhen, China, 201–206.

[15] V. Romero and J. A. Sánchez. 2013. Category-Based Language Models for Handwriting Recognition of Marriage License Books. In *Proc. of ICDAR 2013*. 788–792.

[16] A. Stolcke. 2002. SRILM-an extensible language modeling toolkit.. In *Proc. of the 3rd Interspeech*. 901–904.

[17] J Ignacio Toledo, Manuel Carbonell, Alicia Fornés, and Josep Lladós. 2019. Information extraction from historical handwritten document images with a context-aware neural model. *Pattern Recognition* 86 (2019), 27–36.

[18] A. H. Toselli, A. Juan, D. Keysers, J. González, I. Salvador, H. Ney, E. Vidal, and F. Casacuberta. 2004. Integrated Handwriting Recognition and Interpretation using Finite-State Models. *IJPRAI* 18, 4 (June 2004), 519–539.

[19] E. Vidal and D. Llorens. 1996. Using Knowledge to Improve N-gram Language Modelling Through the MGGI Methodology. In *Proceedings of the 3rd International Colloquium on Grammatical Inference: Learning Syntax from Sentences (ICG! '96)*. Springer-Verlag, London, UK, UK, 179–190. http://dl.acm.org/citation.cfm?id=645516.658257

[20] E. Vidal, F. Thollard, C. De La Higuera, F. Casacuberta, and R.C. Carrasco. 2005. Probabilistic finite-state machines-part II. *IEEE Transactions on PAMI* 27, 7 (2005), 1026–1039.