Information Extraction in Handwritten Marriage Licenses Books Using the MGGI Methodology

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

[1] PRHLT Research Center, Universitat Politécnica de Valéncia, Spain
{vromero,evidal,jandreu}@prhlt.upv.es,
[2] Computer Vision Center, Department of Computer Science,
Universitat Autónoma de Barcelona, Spain. afornes@cvc.uab.es

Abstract. Historical records of daily activities provide intriguing insights into the life of our ancestors, useful for demographic and genealogical research. For example, marriage license books have been used for centuries by ecclesiastical and secular institutions to register marriages. These books follow a simple structure of the text in the records with a evolutionary vocabulary, mainly composed of proper names that change along the time. This distinct vocabulary makes automatic transcription and semantic information extraction difficult tasks. In previous works we studied the use of category-based language models and how a Grammatical Inference technique known as MGGI could improve the accuracy of these tasks. In this work we analyze the main causes of the semantic errors observed in previous results and apply a better implementation of the MGGI technique to solve these problems. Using the resulting language model, transcription and information extraction experiments have been carried out, and the results support our proposed approach.

Keywords: Handwritten Text Recognition, Information extraction, Language modeling, MGGI, Categories-based language model

1 Introduction

Handwritten marriage licenses books [8, 7] have been used for centuries by ecclesiastical and secular institutions to register marriages. The information contained in these historical documents is very interesting for demography studies and genealogical research. Therefore, one of the goals of this kind of documents, rather than to transcribe them perfectly, is to extract their relevant information to allow the users to make use of it through semantic searches. Note that, if the perfect transcript is obtained, then identifying the relevant semantic information would be much easier, but it is not mandatory to obtain the perfect transcript.

The automatic transcription of historical documents is currently based on techniques that have been used in Automatic Speech Recognition, such as Hidden Markov Models (HMM) [10] or hybrid HMM and Artificial Neural Networks (ANN) [2] for representing optical models, and n-gram models for language modeling. This is due, in part, to the problems found by traditional Optical Character Recognition (OCR) techniques to segment the linguistic components of these images like characters, words or sentences automatically. Therefore, holistic approaches, that do not need prior segmentation, are needed [5].

The language model plays a fundamental role in the Handwritten Text Recognition (HTR) process by restricting significantly the search space. Although the training of the optical models is still an incipient research field, significant improvements can be obtained by using better language models. For example, in [9], given the regular structure of marriage licenses documents, the use of a category-based language model [6] to both better representing the regularities in marriage license books and for obtaining the relevant semantic information of each record was presented with encouraging results. In [11], a Grammatical Inference technique known as MGGI [12] was studied to improve the semantic accuracy of the category-based language model obtained in [9]. In MGGI, a-priory knowledge is used to label the words of the training strings in such a way that a simple bigram can be trained from the transformed strings. The knowledge used allows the MGGI to produce a language model which captures important dependencies of the language underlying in the handwritten records considered.

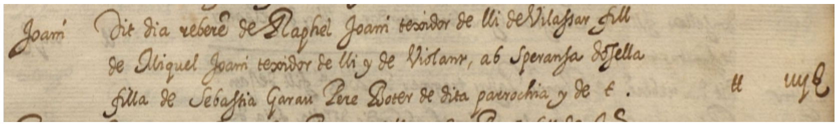
In this paper we analyze the main semantic errors with the category-based language model presented in [11] and relabel the words of the training strings, before applying the MGGI methodology. Our objective is to capture important dependencies of the licenses structure that were not captured in the previous version, such as the relative position of the information within the record.

2 Task description

In this work we have used a handwritten marriage license book from a collection conserved at the Archives of the Cathedral of Barcelona and described in [7]. It is the same book used in previous works such as [9, 11].

Each marriage license typically contains information about the marriage day, groom's and bride's names, the groom's occupation, the groom's and bride's former marital status, and the socio-economic position given by the amount of the fee. This information is not written randomly but the opposite. The groom's information is written first and then the bride's information. Inside the groom's information, the given name and surnames are written first, then the birth town and then the occupation. Then the groom's father information is in a similar order, and then the bride's information. In some cases, additional information is given as well as information about a deceased parent. This structure suggests that 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. 1 shows an example of an isolated marriage license.

As discussed in [11], a problem when transcribing handwritten marriage license books by means of HTR methods is that the classical n-gram language models can be very inaccurate due to the restrictions of the underlying language. Contrary to popular languages such as english or spanish, these documents are written in old catalan, and the amount of available datasets for training in this language are very scarce.



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

Fig. 1. Example of a marriage license.

Another problem is due to the special vocabulary of this collection, since it is mainly composed of proper names. For example, consider the license of the Fig. 1 that starts with the following sentence referred to the groom:

```
Dit dia rebere$ de Raphel Joani texidor de lli de Vilassar ...
```

The translation of this sentence is:

```
That day we received from Raphel Joani linen weaver from Vilassar ...
```

Note that is quite difficult to predict the word `Raphel` from the previous words since any (groom's) given name can appear in this position. Something similar occurs for other words, like `Joani`, (groom's surname) `linen` or `Vilassar` (groom's town). However, if the groom's given name is categorized, the number of contexts in the n-gram model is reduced and, therefore, is easy to predict the correct word. This is the idea described in the following section.

3 Category-based HTR

As shown in [9], the use of a category language model in the handwritten text recognition process can benefit both, the handwritten accuracy and the semantic information extraction process. This improvement is due to two main reason. Firstly, given that category-based language models shares statistics between words of the same category, category-based models are able to generalize to word patterns 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.

In this paper, the same semantic categories defined in [9] 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. Next, the annotated license corresponding to the image in Fig. 1 is shown. Each semantic label (marked into brackets) is immediately after the relevant word:

```
Dit dia rebere$ de Raphel[GrName] Joani[GrSurname] texidor_de_lli[GrOc]
de Vilassar[GrResi] fill de Miquel[GrFaName] Joani[GrFaSurname]
texidor_de_lli[GrFaOc] y de Violant[GrMoName], ab Sperensa[BrName]
do$sella filla de Sebastia_Garau[BrFaName] Pere[BrFaSurname]
Boter[BrFaOc] de dita_parrochia[BrFaResi] y de t.[BrMoName]
```

As shown in the example, only some words had relevant semantic information. Our categorization focuses on these relevant words, and a partially categorized corpus was obtained. Words that do not have a category could be viewed as categories that contain a single word. For instance, we can introduce the category "DIA" containing only the word "dia". On the other hand, a word may belong to several categories. For example, the word *Ferrer* (that could be translated as Smith) could belong to the categories *husband surname*, *husband profession*, *father husband surname*, *father husband profession*, *bride surname*, etc.

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 asociated to the word sequence. Following the discussion presented in [9, 11], from the decoding process, we can obtain not only the best word sequence hypothesis, $\hat{\mathbf{w}}$, but also the best sequence of semantic categories $\hat{\mathbf{c}}$ used in the most probable sentence:

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

$P(\mathbf{x} \mid \mathbf{w})$ represents the optical-lexical knowledge and is typically approximated by concatenated character models, usually HMMs [4]. $P(\mathbf{w} \mid \mathbf{c})$ is the word-category distribution, approximated by an 1-gram for each category. $p(\mathbf{c})$ is the probability of the categories sequence and is approximated by an n-gram.

4 Language Modeling using Morphic Generator Grammatical Inference (MGGI)

As discussed in [11], it is well known that n-gram models are just a subclass of probabilistic finite-state machines (PFSM) [13, 14]. 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 approach (word) string distributions involving the kind of long-span dependencies which are common in natural language.

While learning PFSMs from training strings is in general hard, there is a not-very-well-known 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 the restrictions which are interesting for the task in hand, to ensure that the trained finite-state models will comply with these restrictions. MGGI was introduced in 1987 [3], within the framework of *Grammatical Inference* for Syntactic Pattern Recognition. It is based on the well known "morphism theorem of regular languages [1], which states that every regular language (generated or accepted by a finite-state machine) can be obtained by applying an appropriate word-by-word morphism to the strings of a *local language* over some suitable vocabulary. A probabilistic extension of this theorem is given in [14], where it is also shown that a probabilistic local language is exactly the same language described by a bigram language model.

In MGGI, a-priory knowledge is used to label the words of the training strings in such a way that a simple bigram can be trained from the transformed strings. Then an inverse transformation (the *morphism*) is applied to this bigram to obtain a PFSM which deals with the restrictions conveyed by the initial string transformation [3, 14]. A direct applications of these ideas to build accurate PFSM language models for automatic speech recognition can be seen in [12].

In [11], 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.

In this work, we checked the most frequent errors committed by the language model obtained after the MGGI application in the same way than in [11] and relabel the training samples in such a way that allow to solve the detected errors. One of the most common errors was the mis-categorization of the groom's father information as groom's information. The following example shows an example of this kind of errors, where the groom's father name has been wrongly labeled as the groom's name and the same occurred with the surname and the profession:

```
... fill de Miguel[GrName] Joani[GrSurname] teixidor_de_lli[GrProf] y ...
```

This clearly happened because the bigram *"de [GrName]"* had higher probability than the bigram *"de [GrFaName]"*, since groom's information appears more often than the groom's father information. The same occurs with the bride's mother information and with the bride's information. This suggests that a better generalization of the training text could be achieved by just tagging all the text tokens (categories and words) with labels that help distinguishing their relative position in the record.

In the vast majority of the records that we considered, the information of the groom and his parents is separated by the word *"fill"* ("son" in English). Similarly, the information of the bride and her parents is separated by the word *"filla"* ("daughter" in English). Therefore, it is straightforward to label all the tokens which precede the word *"fill"* with the suffix *"G"*, those appearing between *"fill"* and *"ab"* as *"F"*, those between *"ab"* and *"filla"* as *"B"* and the rest as *"A"*. By applying this labeling scheme to the categorized training transcripts of the license of the Figure 1, the following training text is obtained:

```
DitG diaG rebere$G deG [GrName]G [GrSurname]G [GrProf]G deG [GrResi]G
fillF deF [GrFaName]F [GrFaSurname]F [GrFaProf]F yF deF [GrMoName]F
,F ab [BrName]B do$sellaB fillaA deA [BrFaName]A [BrFaSurname]A
[BrFaProf]A deA [BrFaResi]A yA deA [BrMoName]A
```

Given that the words *"fill"*, *"filla"* and *"ab"* only appear once in each record, the relabeling can be automatically done, so there is no extra manual work by the expert user. After training the category-based bigram, the inverse transformation required by MGGI (the word-by-word morphism) consists in removing these suffixes. The resulting PFSM adequately models the dependencies conveyed by the adopted labeling.

5 Experimental Framework

We have used the ESPOSALLES[3] database [7], a 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. It contains around 60,000 running words from a lexicon of 3,500 different words. A paleographer transcribed and annotated the 40 categories defined by demographers, as described in [9].

Seven partitions of 25 pages were used for cross-validation. The pages were divided into line images, and normalized as explained in [7]. For each line image, we extracted a sequence of feature vectors [10] 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.

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 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 [4]. A category-based bi-gram was estimated using the MGGI methodology from the training set. The words in the test partition that do not appear in the training set, named Out of Vocabulary (OOV) words, were added as singletons to the corresponding word category distribution. The word category distributions were modeled by uni-grams.

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. 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. 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:

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

6 Results

The proposed model has been compared to our initial work on MGGI [11] and our baseline system [9], consisting in a HMM-based HTR system using a category-based 2-gram language model (CB-HTR).

Table 1 presents the results in terms of WER, Precision and Recall. The WER remains the same because the MGGI technique is focused on the semantic labeling. However, the increase in the performance in information extraction is

[3] It is publicly available at: `http://dag.cvc.uab.es/the-esposalles-database/`

significative. In the first case, the mean Precision and Recall are computed for the absolute number of instances. In the second case, the mean Precision and Recall are computed by averaging the Precision and Recall for each one of the categories. As it can be observed, the absolute values are higher because there are some categories that appear in few cases, and consequence, the ability of the model to learn is lower. Also, when analyzing the Precision and Recall of the individual categories, we have observed that, when a category is very frequent (e.g. Groom's surname or occupation), the performance is higher, probably due to the higher amount of training data. In low populated categories (e.g. Bride's Residence), the behavior is just the opposite. For example, there are 1736 instances of the Groom's surname, and the MGGI obtains a Recall of 84'3%. Contrary, there are 46 instances of the Bride's surname, and the performance of the MGGI decreases, obtaining a Recall of 69'7%.

Table 1. Word Error Rate (WER), precision (π) and recall (ρ) obtained with the category-based HTR system (CB) and with the MGGI HTR systems (MGGI). 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 [9]	10.1	79.2	66.6	73.5	65.2
MGGI [11]	10.1	85.3	76.2	78.3	72.2
MGGI (our approach)	10.1	87.8	82.3	80.7	76.2

Finally, it must be noted that we consider an error whenever the semantic category or the transcription are incorrect. Therefore, if a word transcription is incorrect, we will also consider it as a semantic labeling error, no matter if the category is correct. Consequently, the computation of the semantic labeling error is pessimistic, which means that it will never be lower than WER.

7 Conclusions

In this paper, we have improved the MGGI methodology for information extraction and for automatically transcribing a marriage license book. Given the fixed structure of the information included in the license, we have used it to label the words of the training strings. The labels are chosen in such a way that a bigram trained with the labeled strings deals with restrictions that a simple category-based language model can not. We can see that the MGGI methodology can be useful to automatically extract the relevant information, helping the user in this hard task. As a future work, we would like to investigate how to discover the structure in an automatic way.

Acknowledgment

This work has been partially supported through the European Union's H2020 grant READ (Recognition and Enrichment of Archival Documents) (Ref: 674943), the European project ERC-2010-AdG-20100407-269796, the MINECO/FEDER, UE projects TIN2015-70924-C2-1-R and TIN2015-70924-C2-2-R, and the Ramon y Cajal Fellowship RYC-2014-16831.

References

1. S. Eilenberg. *Automata, Languages, and Machines.* Number pt. 1 in Automata, Languages, and Machines. Academic Press, 1974.
2. S. España-Boquera, M. Castro-Bleda, J. Gorbe-Moya, and F. Zamora-Martínez. Improving offline handwriting text recognition with hybrid HMM/ANN models. *IEEE Trans. on PAMI*, 33(4):767–779, 2011.
3. P. Garcia, E. Vidal, and F. Casacuberta. Local languages, the succesor method, and a step towards a general methodology for the inference of regular grammars. *IEEE Transactions on PAMI*, (6):841–845, 1987.
4. F. Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, 1998.
5. U.-V. Marti and H. Bunke. Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System. *IJPRAI*, 15(1):65–90, 2001.
6. T. Niesler and P. Woodland. A variable-length category-based n-gram language model. In *Proc. of ICASSP-96*, volume 1, pages 164 –167 vol. 1, may 1996.
7. V. Romero, A. Fornés, N. Serrano, J. A. Sánchez, A. Toselli, V. Frinken, E. Vidal, and J. Lladós. The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. *Pattern Recognition*, 46:1658–1669, 2013.
8. V. Romero, J.-A. Sánchez, N. Serrano, and E. Vidal. Handwritten text recognition for marriage register books. In *Proc. of the ICDAR 2011*, pages 533–537.
9. V. Romero and J. A. Sánchez. Category-based language models for handwriting recognition of marriage license books. In *Proc. of ICDAR 2013*, pages 788–792, 2013.
10. A. H. Toselli, A. Juan, D. Keysers, J. González, I. Salvador, H. Ney, E. Vidal, and F. Casacuberta. Integrated Handwriting Recognition and Interpretation using Finite-State Models. *IJPRAI*, 18(4):519–539, June 2004.
11. E. V. Verónica Romero, Alicia Fornés and J. A. Sánchez. Using the MGGI methodology for category-based language modeling in handwritten marriage licenses books. In *ICFHR*, Shenzhen, China, 2016.
12. E. Vidal and D. Llorens. 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, pages 179–190, 1996.
13. E. Vidal, F. Thollard, C. De La Higuera, F. Casacuberta, and R. C. Carrasco. Probabilistic finite-state machines-part I. *IEEE Transactions on PAMI*, 27(7):1013–1025, 2005.
14. E. Vidal, F. Thollard, C. De La Higuera, F. Casacuberta, and R. C. Carrasco. Probabilistic finite-state machines-part II. *IEEE Transactions on PAMI*, 27(7):1026–1039, 2005.