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

I. INTRODUCTION

Historical records of daily activities provide intriguing insights into the life of our ancestors, useful for demography studies and genealogical research [1]. Handwritten marriage licenses books [2] are one of these records that have been used for centuries by ecclesiastical and secular institutions to register marriages. The information contained in these historical documents is very interesting for migratory studies, population research and genealogical investigation. Therefore, one of the goals of this kind of documents, rather than to transcribe perfectly the documents, is to extract the 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.

For typical handwritten text images of historical documents, currently available text image recognition technologies are not suitable. Traditional Optical Character Recognition (OCR) is simply not usable since the linguistic components like characters, words or sentences can not be isolated automatically. Therefore holistic approaches that do not need prior segmentation are needed [3]. Thus, HTR of historical documents is currently based on techniques that have been used in Automatic Speech Recognition. In this way, Hidden Markov Models (HMM) [4] or hybrid HMM and Artificial Neural Networks (ANN) [5] are used for representing optical models, and n-gram models for language modeling.

The language model plays a fundamental role in the HTR process, by restricting significantly the search space. For tasks with a vocabulary of medium size current HTR state-of-the-art prototypes provide word error levels that roughly range from 10 to 40% [4], [2]. 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 [6], given the regular structure of marriage licenses documents, the use of a category-based language model [7] for both better representing the regularities in marriage license books and for obtaining the relevant semantic information of each record was studied. In contrast to this, there are works where the semantic interpretation of the recognition output is carried out in a second step using natural language understanding techniques [8].

In this paper, we follow the first approach and analyze the main semantic errors occurred using category-based language models and apply a Grammatical Inference technique known as “Morphic Generator Grammatical Inference” (MGGI) [9] to improve the semantic accuracy of the language model obtained. 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 dependencies of the language underlying in the handwritten records considered.

II. TASK DESCRIPTION

In this paper we used a book from a collection of Spanish marriage license books conserved at the Archives of the Cathedral of Barcelona and described in [2]. Fig. 1 shows a page of marriage licenses from the book used in this paper. 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 infor-
mation 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. 2
shows an example of an isolated marriage license.

A problem when transcribing handwritten marriage li-
cense books by means of HTR methods is that the classical
n-gram language models can be very inaccurate due to the
special vocabulary of the task, which is composed mainly
of proper names (given names, surnames, town names, etc.).
A classical n-gram language model can have difficulties to
predict the probability of a word if the word to be predicted
is a proper name. For example, consider a license that starts
with the following sentence referred to the groom:


As shown in the example, only some words of each
license had relevant semantic information, the proposed
categorization involved classifying only some words in the
vocabulary and not all of them. In this way, a partially
categorized corpus was obtained, that is, not each word
had a category associated to it. Words that had not 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
\text{Ferrer} (that could be translated as Smith) could belong
to the categories husband surname, husband profession,
father husband surname, father husband profession, bride
surname, father bride surname, etc.

Formally speaking, let \( x = x_1 \ x_2 \ ... \ x_m \) be a
handwritten sentence image represented by a feature vector
sequence. The HTR problem is formulated as the problem of

III. CATEGORY-BASED HTR

As shown in [6], the use of a category language model in
the handwritten text recognition process can benefit both, the
handwritten accuracy and the semantic information extrac-
tion 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 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, following the ideas presented in [6], some
categories have been defined taking into account the seman-
tic information included in the licenses: groom’s (\text{Gr})
given name and surname, bride’s (\text{Br}) given name and surname,
parents’ (\text{Fa} and \text{Mo}) given names and surnames, occupa-
tions (\text{Oc}), place of residence (\text{Resi}), geographical origin,
etc. Then, a category-based language model was generated
and integrated into the handwritten text recognition process.
In the next text, the annotated license corresponding to the
image in Figure 2 is shown. Each semantic label (marked
into brackets) is immediately after the relevant word:


The translation of this sentence is

\text{That day we received from Raphel Joani linen weaver

from Vilassar ...}

Note that is quite difficult to predict the word \text{Raphel}
from the previous words since any (groom’s) given name
can appear in this position. Something similar occurs for

other words, like \text{Joani}, (groom’s surname) \text{linen} or
\text{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.
finding the most likely word sequence, \( \mathbf{w} = w_1 \ldots w_n \), i.e., \( \mathbf{w} = \arg \max \mathbb{P}(\mathbf{w} | \mathbf{x}) \). Using the Bayes’ rule we can decompose this probability into two probabilities, \( \mathbb{P}(\mathbf{x} \mid \mathbf{w}) \) and \( \mathbb{P}(\mathbf{w}) \), representing optical-lexical knowledge and syntactic knowledge, respectively:

\[
\hat{\mathbf{w}} = \arg \max \mathbb{P}(\mathbf{w} \mid \mathbf{x}) = \arg \max \mathbb{P}(\mathbf{x} \mid \mathbf{w}) \cdot \mathbb{P}(\mathbf{w})
\]

(1)

\( \mathbb{P}(\mathbf{x} \mid \mathbf{w}) \) is typically approximated by concatenated character models, usually HMMs [10], while \( \mathbb{P}(\mathbf{w}) \) is approximated by a language model, in this work we use a category-based language model [6].

By further considering the sequence of semantic categories, \( \mathbf{c} = c_1 \ldots c_l \), associated to the word sequence as a hidden variable in equation (1), approximating the sum by the dominating term, and following the same assumptions presented in [6], we can rewrite previous equation as:

\[
\hat{\mathbf{w}} = \arg \max \mathbb{P}(\mathbf{x} \mid \mathbf{w}) \sum \mathbb{P}(\mathbf{c}, \mathbf{w})
\]

(2)

\[
\approx \arg \max \mathbb{P}(\mathbf{x} \mid \mathbf{w}) \max \mathbb{P}(\mathbf{c}) \cdot \mathbb{P}(\mathbf{w} \mid \mathbf{c})
\]

(3)

\[
\approx \arg \max \mathbb{P}(\mathbf{x} \mid \mathbf{w}) \max \mathbb{P}(\mathbf{c}) \prod_{i=1}^{l} \mathbb{P}(c_i \mid c_{i-1} \ldots c_{i-n+1}) \cdot \mathbb{P}(w_i \mid c_i)
\]

(4)

where \( \mathbb{P}(w_i \mid c_i) \) is computed from the word-category distribution and \( \mathbb{P}(c_i \mid c_{i-1} \ldots c_{i-n+1}) \) is computed from an \( n \)-gram of categories.

Finally, as explained in [11], from the decoding process, we can obtain not only the best word sequence hypothesis, but also the best sequence of semantic categories used in the most probable sentence:

\[
(\hat{\mathbf{c}}, \hat{\mathbf{w}}) \approx \arg \max \mathbb{P}(\mathbf{x} \mid \mathbf{w}) \cdot \mathbb{P}(\mathbf{w} \mid \mathbf{c}) \cdot \mathbb{P}(\mathbf{c})
\]

(5)

IV. LANGUAGE MODELING USING MGGI

It is well known that \( n \)-gram models are just a subclass of probabilistic finite-state machines (PFSM) [12], [13]. 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. 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 \( ab^i c \) or \( db^i e \) and is low or null for other strings such as \( ab^i e \) and \( db^i c \), where \( i \) is any arbitrarily large integer. However, such a distribution can be exactly modeled by a very simple PFSM (see [13], Sec. 2.1.3).

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. For instance, in the previous example, one can rely on the known legal starting and ending words to obtain a PFSM which accurately approaches the distribution aimed at (see [13], Sec. 2.2).

MGGI was introduced in 1987 [14], within the framework of Grammatical Inference for Syntactic Pattern Recognition. It is based on the well known “morphism theorem of regular languages” [15], 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 [13], where it is also shown that a probabilistic local language is exactly the same as a bigram language model.

In MGGI, a-priori 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 [14], [13]. A direct application of these ideas to build accurate PFSM language models for automatic speech recognition can be seen in [9].

In this work, rather than using a plain bigram trained on raw word strings, a category-based bigram is used for the application of MGGI. This way, the resulting PFSM will additionally provide the helpful generalizations entailed by word categorization. The knowledge used to define and label relevant word categories is expanded with additional word labels which allows the MGGI to produce a PFSM which captures important dependencies of the language underlying in the handwritten records considered.

To this end, we checked the most frequent errors committed by a standard category-based bigram, such as the one used in [6]. One of the most common errors, clearly due to a wrong bigram generalization, was the mis-categorization of...
the bride’s family information as groom’s information. The following example shows an example of this kind of errors, where the bride’s father name has been wrongly labeled as the groom’s father name and the same occurred with the surname and the profession:

\[
\ldots ab \text{Sperensa}\{BrFaName\} do\$se\text{lla } \text{filla de}
\text{Sebastia} \text{Garau}\{GrFaName\} \text{Pere}\{GrFaSurname\}
\text{Boter}\{GrFaOc\} \text{de } \text{dita_parrochia}\{BiFaResi\} \text{y}
\ldots
\]

This clearly happened because the bigram “de \{GrFaName\}” had higher probability than the bigram “de \{BrFaName\}”, since groom’s family information appears more often than that bride’s family 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 considered, the groom’s and bride’s information are separated by the word “ab” (“with” in English). Therefore, it is straightforward to label all the tokens which precede the word “ab” with the suffix “G” and those appearing after “ab” with “B” (meaning that the informations correspond to the Groom and the Bride, respectively). By applying this labeling scheme to the categorized training transcripts of the license of the Figure 2, the following training text is obtained:

\[
\text{Dit}\text{G diaC rebere$G \text{ deG}\{GrFaName\}\text{G}\{GrSurname\}\text{G}}
\text{GrProf}\text{C deG}\{GrResi\}\text{C fillG deG}\{GrFaName\}\text{G}
\text{GrFaSurname}\text{G}\{GrFaProf\}\text{G yG deG}\{GrMoName\}\text{G}
\text{G ab \{BrName\}B do\$se\text{llaB } \text{fillaB deB}
\text{BrFaName}\text{B}\{BrFaSurname\}\text{B}\{BrFaProf\}\text{B deB}
\text{BrFaResi}\text{B yB deB}\{BrMoName\}\text{B}
\]

After training a category-based bigram, the inverse transformation required by MGGI (the word-by-word morphism) consists just in removing these suffixes “G” and “B”. The resulting PFSM adequately models the dependencies conveyed by the labeling adopted.

V. EXPERIMENTAL FRAMEWORK

To assess how using the MGGI for category-based language modeling can benefit the handwriting recognition and the semantic information extraction, different experiments were carried out. The corpus, the assessment measures and the obtained results are explained next.

A. Corpus

The experiments were performed on the publicly available ESPOSALLES\(^1\) database [2], which was compiled from a marriage license book conserved at the Archives of the Cathedral of Barcelona. We used the version labeled with the semantic information presented in [6].

The corpus was written in old Catalan by only one person between 1617 and 1619. It is composed by 173 pages that contain 5,447 lines grouped in 1,747 licenses. The whole manuscript was transcribed line by line by an expert paleographer. The complete annotation contains around 60,000 running words from a lexicon of around 3,500 different words. The database was also labeled with the semantic information of the licenses. 40 different categories were defined by demographer experts and the relevant words in each license were manually labeled with the corresponding category as shown in section III.

The standard partition proposed in [2], consisting of seven consecutive blocks of 25 pages has been used in the experiments. Table I shows the average values of the statistics related with the different partitions.

<table>
<thead>
<tr>
<th>Number of:</th>
<th>Total</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages</td>
<td>173</td>
<td>24.7</td>
</tr>
<tr>
<td>Licenses</td>
<td>1,747</td>
<td>249.6</td>
</tr>
<tr>
<td>Lines</td>
<td>5,447</td>
<td>778.1</td>
</tr>
<tr>
<td>Run. words</td>
<td>60,777</td>
<td>8682.4</td>
</tr>
<tr>
<td>OOV</td>
<td>–</td>
<td>361</td>
</tr>
<tr>
<td>Lexicon</td>
<td>3465</td>
<td>1070</td>
</tr>
<tr>
<td>Semantic labels</td>
<td>21386</td>
<td>3055.1</td>
</tr>
</tbody>
</table>

B. System setup

The seven different partitions were used in these experiments for cross-validation. That is, 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 as explained in [2]. Then, appropriate filtering methods were applied to remove noise, improve the quality of the image and to make the documents more legible. Afterwards, the skew and the slant of each line were corrected. Finally the size was normalized separately for each line.

Each preprocessed line image was represented as a sequence of feature vectors. In this work we used the features described in [4] based on the gray level of the image. As explained in [2], the ESPOSALLES database is provided at two different levels: line level and license level. In this work we have carried out experiments at license level. Given that the lines belonging to each license was known, feature sequences extracted from the lines could be easily merged into whole license line images.

The characters were modeled by continuous density left-to-right HMMs with 6 states and 64 Gaussian mixture components per state. These models were estimated from training text images represented as feature vector sequences using the Baum-Welch algorithm.

A category-based bi-gram was estimated using the MGGI methodology from the training transcriptions of the text line

\(^1\)It is publicly available at: http://www.cvc.uab.es/5cofm/groundtruth
images. The out of vocabulary (OOV) words (words of the test partition that do not appear in the training partition) belonging to a category seen in training were added as singletons to the corresponding word category distribution. For OOV words that belong to a category that has not been seen in training, 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. The decoding was carried out by the Viterbi algorithm [10].

C. Assessment Measures

Different evaluation measures were adopted to assess the HTR and the relevant information extraction performance. The quality of the transcription is given by the well known Word Error Rate (WER). It is 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.

On the other hand, to assess the quality of the information extraction performance we have used the standard precision and recall measures. We define precision and recall in terms of the number of relevant words in the dataset and the number of relevant words retrieved by the system. Relevant words are those words that belong to one of the 40 relevant categories defined by demographer experts. For instance, in the example shown in previous sections, the relevant words are those associated to a category: Raphel, Joani, texidor_de_Jii, Villassar, Miquel, ... Let $R$ be the number of relevant words contained in the document, let $D$ be the number of relevant words that the system has detected, and let $C$ be the number of the relevant words correctly detected by the system. Precision ($\pi$) and recall ($\rho$) are computed as:

$$\pi = \frac{C}{D}, \quad \rho = \frac{C}{R}$$

Finally, it must be said that we consider an error whenever the semantic category or the transcription are incorrect. This means that if a word transcription is incorrect, then we will consider it also a semantic labeling error, although its category is correct. Consequently, the computation of the semantic labeling error is pessimistic, which means that it will never be lower than WER.

VI. RESULTS

Our proposed model has been compared to a baseline system proposed in [6], which consists in a HMM-based HTR system using a category-based 2-gram language model (CB-HTR). Table II presents the experimental results in terms of WER, Precision and Recall. Although the WER remains the same because the MGGI technique is focused on the semantic labeling, the performance in information extraction significantly improves. 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.

<table>
<thead>
<tr>
<th></th>
<th>WER</th>
<th>$I_{\pi}$</th>
<th>$I_{\rho}$</th>
<th>$C_{\pi}$</th>
<th>$C_{\rho}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB</td>
<td>10.1</td>
<td>79.2</td>
<td>66.6</td>
<td>73.5</td>
<td>65.2</td>
</tr>
<tr>
<td>MGGI</td>
<td>10.1</td>
<td>85.3</td>
<td>76.2</td>
<td>78.3</td>
<td>72.2</td>
</tr>
</tbody>
</table>

Table II shows the detailed results for some categories. It is worth to notice that, in both methods, $\pi$ and $\rho$ are usually high in very populated categories (e.g. Groom’s and Bride’s names), and tend to decrease in categories with very few instances (e.g. Bride’s residence and origin). This behavior is probably due to the few training data for these low populated categories.

Concerning the comparisons, the first observation is that $\pi$ and $\rho$ are usually higher when using MGGI rather than CB. In many categories, usually in those corresponding to the bride information (e.g. Father’s Bride Name and Father’s Bride Surname) the improvement is over 20 points, whereas in others (e.g. Father’s Groom Profession, Groom’s Origin) the performance is similar. In fact, the CB methodology outperforms the MGGI only when the number of instances is lower (e.g. Bride’s residence and origin). Secondly, the absolute number of repeated categories $\text{rep}$ that are found in the same register decreases when using the MGGI technology. In this sense, whenever the same category is found several times in one register, then only the first one is taken into account.

VII. CONCLUSIONS

In this paper, we have studied the use of the MGGI methodology for category-based language modeling to relevant information extraction and for automatically transcription of 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. From the results we can see that using the MGGI methodology can be useful to automatically extract the relevant information, helping the user in this hard task.

Finally, given that the MGGI requires a-priory knowledge of the task that it is not always available, as future work, we intend to compute the anchor points to apply the new
labelling required in the MGGI automatically. This can be done using combination techniques based in confusion networks, such as that presented in [16].

ACKNOWLEDGMENT

This work has been partially supported through the European Union’s H2020 grant READ (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


