System Description of CITlab’s Recognition & Retrieval Engine for ICDAR2017 Competition on Information Extraction in Historical Handwritten Records

Tobias Strauß · Max Weidemann · Johannes Michael · Gundram Leifert · Tobias Grüning · Roger Labahn

Received: date / Accepted: date

Abstract We present a recognition and retrieval system for the ICDAR2017 Competition on Information Extraction in Historical Handwritten Records which successfully infers person names and other data from marriage records. The system extracts information from the line images with a high accuracy and outperforms the baseline. The optical model is based on Neural Networks. To infer the desired information, regular expressions are used to describe the set of feasible words sequences.

Keywords Text recognition, information retrieval, regular expressions, recurrent neural networks

1 Introduction

There is a huge amount of handwritten texts containing information of past times which are valuable but not yet accessible. The ICDAR2017 Competition on Information Extraction in Historical Handwritten Records encourages research in the field of automatic retrieval systems by providing training data from marriage records.

We present a bottom-up approach which processes the writing resulting in a matrix of probabilities per character and position. A two step process finds the most likely character sequence according to this matrix and previously defined regular expressions covering the expected structure and assigns information containing parts of this sequence to the specific categories.

2 Task

The data set consists of well-written marriage records of the 17th century from the Esposalles database. The task is to extract words of categories of interest like name, surname, location and state (Track 1) and assign them to persons like husband, wife, husband’s father, wife’s mother etc. (Track 2) from the given line images. A sample record is given in Fig. 1.

The organizers provided 970 records (consisting of 3070 lines) for training and validation including transcriptions, categories and person classes. The test set comprises 757 lines from 253 records. The major problem with the data set is to parse the variations of the language. Promising sequence 2 sequence approaches (see [Sutskever et al., 2014]) could solve this issue in the future without manual effort which is still necessary for the proposed system.

3 Recognition Engine and Retrieval

3.1 Preprocessing

Given the line polygon, we apply certain standard pre-processing routines, i.e.

- image normalization: contrast enhancement (no binarization), size;
- writing normalization: line bends, line skew, script slant.

Tobias Strauß · Max Weidemann · Johannes Michael · Gundram Leifert · Roger Labahn
Institute of Mathematics
University of Rostock
18051 Rostock
Tel.: +49-381-4986633
E-mail: {tobias.strauss, max.weidemann, johannes.michael, gundram.leifert, roger.labahn}@uni-rostock.de

Tobias Grüning
Planet AI
Warnowufer 60
18057 Rostock
E-mail: tobias.gruening@planet.de
Then, images are further unified by CITlab’s proprietary writing normalization: The writing’s main body is placed in the center part of an image of fixed 96px height. While the length-height ratio of the main body stays untouched, the ascenders and descenders are squashed to focus the network’s attention on the more informative main body.

3.2 Neural Network

The preprocessed images are fed into a neural network of the architecture described in Table 1. The implementation is based on TensorFlow (see [Abadi et al., 2016]). The three convolutional layers additionally apply batch normalization (see [Ioffe and Szegedy, 2015]) before and local response normalization (see [Krizhevsky et al., 2012]) after applying the ReLU activation function. The BLSTM layers are trained with dropout (applied to the output and keep ratio of 0.5, see [Gal and Ghahramani, 2016]).

| conv | conv | BLSTM | conv | BLSTM | fully |
|------|------|-------|------|-------|-------|
| Neurons | 8 | 32 | 256 | 64 | 512 | 62 |
| stride | 4x3 | 4x3 | 1x2 |

The last layer is the fully-connected layer and contains 62 neurons. One of these neurons represents a garbage label ◯ (not-a-character or NaC in the following) and the others correspond to the 61 characters appearing in the ground truth. We denote the character set of the 61 characters by $\mathcal{A}$ and label set $\mathcal{A} \cup \{\emptyset\}$ by $\mathcal{A'}$ here and after. The loss function is the typical CTC-loss (see [Graves et al., 2006]). The network is trained 150 epochs by RMSProp (see [Tieleman and Hinton, 2012]) where one epoch contains 4096 randomly sampled line images. The initial learning rate is 0.002 and decayed after every third epoch by a factor of 0.95.

The output of the last layer is softmax transformed such that the output of the neural network is a matrix $Y \in [0,1]^{T \times 62}$ of variable length $T$. For each row $t$, $\sum_{l \in \mathcal{A'}} y_{t,l} = 1$. We call $Y$ ConfMat.

3.3 Decoding

Certain lines (and thus the corresponding ConfMats) belong to the same record. These ConfMats are concatenated to one whole ConfMat per record. The encoded text follows specific rules which can be formulated as regular expressions. To decode the most likely character sequence according to a regular expression, we use the method described in [Strauß, 2016].

Let $\mathcal{F} : \mathcal{A'} \rightarrow \mathcal{A}$ be the mapping which deletes consecutive identical labels and removes all NaCs, e.g. $\mathcal{F}(\emptyset a \emptyset ab) = \mathcal{F}(\emptyset a \emptyset aab) = aab$. The probability of a label sequence $l$ given a line image $X$ is calculated by $P(l \mid X) = \prod_{t=1}^{T} y_{t,l}$, if the ConfMat and the label sequence are both of length $T$ and 0 otherwise. The most likely character sequence $z$ maximizes
\[
\sum_{l \in \mathcal{F}^{-1}(z)} P(l \mid X). \quad \text{Since there is typically one dominant label sequence, we substitute the sum by the maximum:} \\
\begin{align*}
    z^* := \arg\max_z & \quad \max_{l \in \mathcal{F}^{-1}(z)} \ P(l \mid X) \\
\end{align*}
\]

The proposed method is based on two steps: A first coarse labeling is done by a regular expression which splits the whole record ConfMat into regions corresponding to the various persons: husband, wife and their parents. The regular expression is generated manually and includes none of the given vocabularies. The structure of the expression is simple: the regions are identified by several keywords which are followed by a region corresponding to a specific person.

The second step processes these regions corresponding to a specific person separately (see Figure 2). Here, the task is to identify names, locations etc. Incorporating a vocabulary yields more reliable transcriptions than using the most likely network output directly. Thus, we include the provided vocabularies into the regular expression. Only the general category vocabularies are used ignoring e.g. those corresponding to specific persons. Even the surname vocabulary alone comprises more than 1200 names such that a beam search is required to decode the most likely character sequence.

The neural network does not model the prior probability \( P(z) \) of a word \( z \) correctly. A simple application of Bayes law (see [Strauß, 2016]) yields a corrected probability \( P_T(z \mid X) \) of the character sequence \( z \) given the image \( X \)

\[
P_T(z \mid X) \propto \frac{P_T(z)}{P_S(z)} P_S(z \mid X)
\]

up to a normalization which is the same for any character sequence given the same image \( X \). Here, \( P_S(z \mid X) \) represent the probability of the neural network as defined above. \( P_S(z) \) is the prior probability implicitly learned by the neural network. This term cannot be measured directly and has to be estimated. The term \( P_T(z) \) is the true (or at least better) prior probability of the character sequence \( z \).

In the competition, the decoded character sequence maximizes

\[
P_T(z) P_S(z \mid X).
\]

That means, \( P_S(z) \) is assumed to be uniformly distributed over \( \mathcal{A}^T \) (which is not true). For \( z = (z_1, \ldots, z_n) \), \( P_T(z) \) is approximated by the product of the relative frequencies of its subwords from the corresponding vocabularies, i.e., \( \prod_i P_T(z_i) \) (if we ignore spaces and words that are not from vocabularies). Any conditional dependency (e.g. the probability of a location or occupation after the surname) is ignored.

To allow also out-of-vocabulary words, we added the most likely characters per position instead of first name or surname. The prior for such an out-of-vocabulary word is a combination of a character probability and a word probability which is negligible small compared to the relative frequency of any vocabulary word.

4 Competition results

We briefly report the results of the complete track. Details can be found in [Fornés et al., 2017]. The score of the ICDAR2017 Competition on Information Extraction in Historical Handwritten Records is equal to the character accuracy if the category and person (basic track: only category) are correct and 0 otherwise.

Besides the baseline and our systems, there is no other submission at line level. Another track of the same competition provides a word segmentation instead of the line as whole image. Task and score are the same for both levels. The best retrieval system at word level performs slightly better than our best system (overall score of 91.97 against 91.56).

Discussion

In Table 2, the competition results are presented (as given in the article of the organizers [Fornés et al., 2017]). We find systematical gaps e.g. the recognition of the name of any person is always more reliable than the surname. The organizers explained this by the greater variability of surnames.

In total the scale of the results are similar except for the categories husband’s mother’s name, other person’s state, wife’s surname and wife’s location. The first two categories are not considered by our expression and also the other competition participants returned 0 scores. This indicates that these categories are rarely presented in the training data and validation data. For the latter
Table 2 Competitition score (based on CER) for CITlab’s recognition and retrieval system on the track complete.

| husband | husband’s father | husband’s mother | other person |
|---------|------------------|------------------|--------------|
| name    | surname | state | location | occupation | name    | surname | state | occupation | name      | surname | state | occupation |
| 96.10   | 88.85   | 92.42  | 90.42    | 88.49       | 94.28   | 86.57  | 78.61  | 92.07       | 96.17     | 0       | 93.93 | 88.06  |

| wife    | wife’s father | wife’s mother |
|---------|---------------|---------------|
| name    | surname | location | occupation | name | surname | location | occupation |
| 98.49   | 36.57   | 97.13   | 66.73   | 91.43 | 94.42   | 87.43   | 89.29   | 89.17 | 95.90 | -      |

two categories the regular expression seems to fit not very well.

5 Conclusion and Outlook

We presented a retrieval algorithm for the ICDAR2017 Competition on Information Extraction in Historical Handwritten Records. The task is to extract information of the various persons from the lines. The proposed system is based on deep recurrent neural networks. Regular expressions are defined to decode the output. The system is able to infer most of the categories with high precision.

A drawback of the proposed system is the relatively high manual effort to define the precise regular expression. In the future, we will work on reducing this effort either by learning the regular expression automatically or applying the powerful seq2seq models which have shown to cope with such kind of tasks.

Acknowledgment

This work was partially funded by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 674943 (READ – Recognition and Enrichment of Archival Documents).

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

Abadi et al., 2016. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. In OSDI, volume 16, pages 265–283.

Fornés et al., 2017. Fornés, A., Romero, V., Baró, A., Toledo, J. I., Sánchez, J. A., Vidal, E., and Lladóš, J. (2017). Icdar2017 competition on information extraction in historical handwritten records. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pages 1389–1394. IEEE.

Gal and Ghahramani, 2016. Gal, Y. and Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. In Advances in neural information processing systems, pages 1019–1027.

Graves et al., 2006. Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning, pages 369–376. ACM.

Ioffe and Szegedy, 2015. Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR, abs/1502.03167.

Krizhevsky et al., 2012. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105.

Strauß, 2016. Strauß, T. (2016). Decoding the Output of Neural Networks: A Discriminative Approach. Doctoral dissertation, Universität Rostock.

Sutskever et al., 2014. Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112.

Tieleman and Hinton, 2012. Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning, 4(2):26–31.