Handwritten Text Recognition Error Rate Reduction in Historical Documents using Naive Transcribers

Vincent Christlein[1], Anguelos Nicolaou[1], Thorsten Schlauwitz[2], Sabrina Späth[2], Klaus Herbers[2], Andreas Maier[1]

Pattern Recognition Lab, Friedrich-Alexander University Erlangen-Nürnberg[1]
Chair for Medieval History, Friedrich-Alexander University Erlangen-Nürnberg[2]

Abstract

Handwritten text recognition (HTR) is a difficult research problem. In particular for historical documents, this task is hard as handwriting style, orthography, and text quality pose significant challenges. Creation of a single multi-purpose HTR system seems to be out of reach for current state-of-the-art systems. Therefore, we are interested in fast creation of specialized HTR systems for a particular set of historical documents. Manual annotation by historical experts is expensive and can often not be applied at a large scale. Instead, we use the transcripts of naive transcribers that may still contain a significant amount of errors. In this paper, we propose to fuse the recognized word-chain with naive transcribers that can be obtained in a cost-effective way. For the actual fusion, we rely on a word-level approach, the so-called Recognizer Output Voting Error Reduction (ROVER). Results indicate that we are able to reduce the Word Error Rate (WER) of an HTR system trained with only few pages from 24.6 % to 19.2% with two additional transcribers with 25.1% and 27.1% WER each. This performance is already close to current state-of-the-art systems trained with significantly more data.

1 Introduction

Hand-written text recognition (HTR) is an active field of research. As it has been investigated for many decades, progress has been remarkable, yet it is still considered an open problem as performance is often considered insufficient for many applications. In particular, the field of historical documents shows a richness of variation. Not only the handwriting itself expresses a large degree of variation, also orthography is not standardized, and in many cases, the scribe varies.

Veröffentlicht durch die Gesellschaft für Informatik e.V. 2018 in
M. Burghardt & C. Müller-Birn (Hrsg.): INF-DH 2018 – Workshopband, 25. Sept. 2018, Berlin
Copyright © 2018 bei den Autoren. https://doi.org/10.18420/infdh2018-13

In order to avoid full HTR, many people rely on word-spotting instead. It was initially applied in speech recognition (Rohlicek et al., 1989) and became also popular for word-image recognition (Rath and Manmatha, 2003). Today, it achieves high performance by approaching text recognition using word image classification and retrieval. In 2014, the idea of joined vector embeddings between text images and pyramids of symbol occurrences improved the state-of-the-art for word spotting (Almazán et al., 2014) and in 2016 Sudholt et al. further expanded the state-of-the-art performance using a deep regression neural network learning the same embeddings (Sudholt et al., 2016). The weakness of all word-spotting approaches is that all solutions are restricted to known vocabularies. This limits the technology to be only applicable to certain applications.

The introduction of the Connectionist Temporal Classification (CTC) loss for training Long Short Term Memory (LSTM) recurrent networks, produced good results and gradually shifted focus from word-spotting to unrestricted large vocabulary recognition (Graves et al., 2009). LSTM networks, among other advantages have demonstrated that they can model more effectively long-range interactions and have shown promise in scaling the problem from recognizing words to recognizing sentences (Bluche, 2016). Although CTC addresses successfully the problem of variable length outputs, encoder-decoder networks (Cho et al., 2014) allow a simpler approach to generating transcriptions. In the most recent competition on the topic in 2017 (Sánchez et al., 2017), Word Error Rates (WER) in the range of 15% could be achieved. Essentially, all participating methods in all tracks were preferring forms of LSTM networks or combinations with Convolutional Neural Networks (CNN) over pure CNNs and other representations. This is a good indication of the dominance of LSTM networks and the progress that open vocabulary text recognition has made. Yet, to train models of sufficient quality, large amounts of accurately annotated training data are required and HTR systems typically are suited only for the problem domain for which they have been developed.

There are many methods reported in literature to fuse the output of several information sources to improve recognition. One of the earliest approaches initially designed for speech recognition is presented in (Fiscus, 1997). The outputs of several recognizers are aligned and a majority vote is formed. An improved version exists (Hillard et al., 2007; Maier et al., 2005), which is for example used for classification tasks. Romero suggests to use several approaches to improve HTR, such as crowd-annotation, speech recognition, and interactivity (Romero, 2017).

In this paper, we propose the use of naive transcribers. To the best of our knowledge, fusion of HTR with naive transcribers has not been reported in literature. In the remainder of this paper, we want to investigate an approach using cheap transcriptions to improve the automatic hand-written text recognition output. A Vietnamese company was hired to transcribe ten historical book pages written in German. We assume that the transcribers were actually not able to understand historical German. Thus, we believe that transcriptions obtained by students or crowd sourcing could be valid alternatives, which have the potential to improve the recognition further.

2 Material and Methods

2.1 Dataset

The dataset consists of pages from the first volume of *Nuremberg letters of correspondence* (Pitz, 1959). This serial source, which ran continually from 1404 to 1738, can be found in the Nuremberg State Archives. In more than 350 volumes, with about 250 folio each, these books contain the outgoing letters of an important actor on the stage of the Holy Roman Empire of the German Nation: the imperial city of Nuremberg. In the highly decentralized empire, the ruling patricians of the trading city were always interested in the latest information and peaceful and balanced power relations, which were threatened by aggressive neighbors and Nuremberg's dependence on safe trade routes. The resulting precise observations of those events make Nuremberg's letters of correspondence a first-rate historical source. And even though a complete analysis of these books has not yet been possible due to their enormous scope, the time required to create a transcription can be considerably accelerated by the means of handwriting recognition and the procedure proposed here. This is facilitated by rare changes in the typeface because a single writer typically wrote the letters for several years. The entries in the letters of correspondence seem to be the concepts of the actual letters, as there are numerous corrections in the form of strikethroughs as well as interlinear and marginal notes.

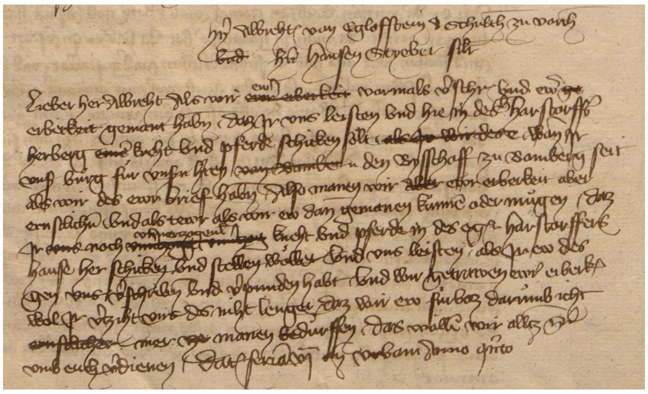

Figure 1 Excerpt of page 56 of the first volume of the Nuremberg letters of correspondence. Image courtesy of State Archive Nuremberg

An example can be seen in Figure 1. The volume is mainly written by a single person. In total, we used 113 pages from the first volume containing 31 087 words. Ground-truth transcription was performed by an expert in medieval sciences, specialized on documents of the Nuremberg area. The data was split into a training set of 103 pages and a test set of 10 pages.

2.2 Handwritten Text Recognition

First, we create a computer-based transcription, which is later fused with the human-based ones. Therefore, we make use of the software Transkribus[1], which is using a state-of-the-art HTR system based on Deep Learning (Leifert et al., 2016; Strauss et al., 2018). The system was trained using a recurrent neural network with CTC loss. In particular, the network consists of six layers, with three convolutional layers, two bidirectional long-short-term memory layers and a fully-connected (FC) layer (Strauss et al., 2018). The FC layer contains 62 neurons, which outputs the probability for the set of 61 characters and a negative class, that means, a non-character class. The network takes whole line images as input and outputs a confidence matrix, which holds the character probabilities at specific line positions. A vocabulary has been trained in advance from the training set, which improves the recognition rate.

2.3 Naive Transcription

We hired a company (Digi-Texx, Vietnam) to transcribe the ten test pages of our historical handwritings by two individuals, that means we obtained two versions of the same text. The company was already hired in the past for transcribing historical handwritten documents in the READ project[2]. The company advertises a character error rate below 5%. However, we found that their performance was lower given our difficult dataset. Scans of the pages were submitted digitally to the company and we obtained the transcripts approximately two weeks after the submission. This process resulted in line-aligned transcripts.

[1] https://transkribus.eu/

[2] https://read.transkribus.eu/

2.4 Word-level Fusion

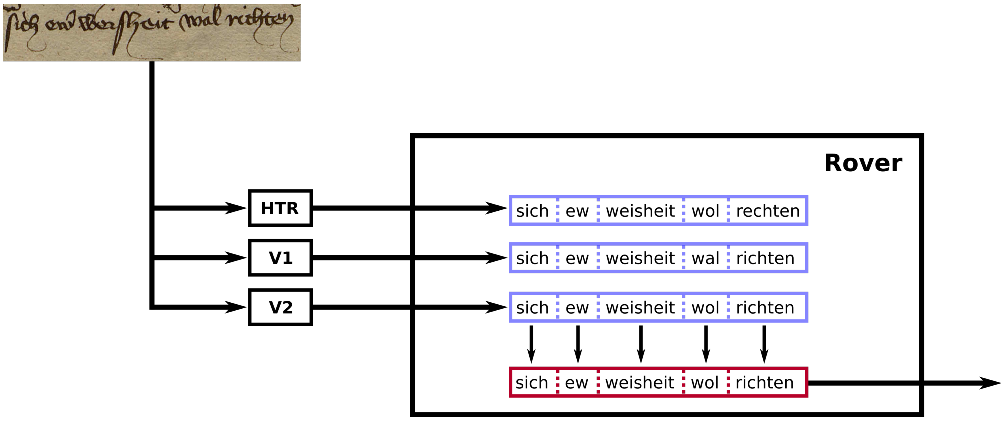
Rover
HTR
V1
V2
sich ew weisheit wol rechten
sich ew weisheit wal richten
sich ew weisheit wol richten
sich ew weisheit wol richten

Figure 2 ROVER aligns the input texts and applies a majority vote for each word position to determine the final word chain.

Fusion was performed using the Recognizer Output Voting Error Reduction (ROVER) method (Fiscus, 1997) implemented in NIST's Speech Recognition Scoring Toolkit (SCTK)[3]. ROVER takes multiple word chains as input and aligns them. Here, line-wise alignment was performed, as the naive transcripts as well as the HTR output resulted in line-resolved output. We chose the HTR system as the reference chain. Next, both naive transcripts were aligned with this reference chain using dynamic programming. As no confidence is known for the naive transcripts, fusion was based on word-level majority voting. The final word chain was then determined as the most frequent word at each position. If there was no clear majority, the reference chain was kept. Figure 2 displays this process schematically.

3 Evaluation

For training, we used 103 pages containing 3144 lines and 28596 words in total. The network is trained with a learning rate of 0.001 for 150 epochs, where 1000 lines correspond to one epoch. The test set comprised ten pages not part of the training set consisting of 323 lines and in total 2491 words. The evaluation is performed using SCTK. Note that we omitted any punctuations due to their ambiguity. While this step naturally improves the WER, we obtain the same relative results.

[3] https://www.nist.gov/itl/iad/mig/tools

3.1 Quantitative Results

Table 1 Single system results and the results obtained by ROVER. All numbers are given in percent.

System	Average Page WER (std. dev.)	Overall WER
HTR	23.1 (±7.0)	24.6
V1	24.9 (±4.7)	25.7
V2	26.4 (±4.3)	27.1
Rover	18.2 (±5.0)	19.2

3.2 Quantitative Results

Table 1 shows the performance of the HTR and the naive transcribers V1 and V2. Interestingly, the overall WER of the HTR system achieves good results of about 25 percent. This is still much lower than the results obtained by recent methods on the ICDAR 2017 HTR competition (Sánchez et al., 2017). However, we neither conducted an exhaustive parameter search to tune the network performance, nor use any specific language models to improve the recognition further. Since our dataset is also quite challenging, we believe that this result represents well the current status of HTR. Given a cleaner dataset, the recognition rates will be significantly lower. The system also outperforms the two naive transcribers. The higher standard deviation of the average page WER of the HTR system suggests that there is more deviation in the results. In fact, there are paragraphs that are very clean while others seem to be written more in haste containing more strikeouts and inter-linear glosses. When fusing the three transcriptions with ROVER, the error rate drops by about more than 5% WER to about 19%.

3.3 Qualitative Results

Table 2 Qualitative results of three random sentences (of page 6), highlighted in blue/red are the improvements/deteriorations by ROVER

GT	klag hie czischn und mich sant Michels tag nu schierst aufschiebn
HTR	klag hieczwischn und und sant Michels tag zu schierst aufschiebn
V1	klag hieczwischn und wies sant wichels tag am schierst aufschriebn
V2	klag hiezwischn und wie sant michels tan nu schierst ausschribn
ROVER	klag hieczwischn und und sant Michels tag zu schierst aufschiebn
GT	wöllen Also ob daz ist daz Ir den vorgen vnsn burgern
HTR	wölle Also ob daz ist daz Ir den vorgen vnsn burgerin
V1	wöllen Also ob daz ist daz Ir den vorgen vnsn burgem
V2	wöllen Also ab daz ist daz Ir den vorgen vnsn burgen
ROVER	wöllen Also ob daz ist daz Ir den vorgen vnsn burgerin
GT	des rechten von de egen hawgen in der zeite helffet So
HTR	des rechten von de egen hawgen in der zeite helffer So
V1	des rechten van de egen uhawgen in der zeite helffet So

V2	des rechten von de egen uhawgen in der zeite helffet So
ROVER	des rechten von de egen uhawgen in der zeite helffet So

Table 2 shows a qualitative result of the different transcriptions. While the transcription of the first line is a good example of the difference between the two transcribers, there is no benefit in using ROVER. Conversely, both humans transcribed correctly the first word of the second line. In the last transcribed sentence, ROVER also successfully corrected one word. Yet, a mistake of both human transcribers can also deteriorate the result as shown in the red-highlighted word. We can conclude that human efforts compensate to some extent the machine errors but also introduces new errors. Maybe more transcriptions, e.g. obtained from another HTR or more naive transcribers, could alleviate this problem.

4 Conclusion

Our first attempt to boost HTR with the help of naive transcribers was successful. The error rate could be reduced from 24.6% to 19.2%, although the naive transcribers only had error rates that were even worse than the HTR system. As such, we believe that the use of naive transcribers is a powerful method to improve HTR performance in a cost-effective way. While the company is used in transcribing historical data, they are no experts in reading medieval text sources. Thus, we believe that the acquisition of transcripts using crowd-source approaches or by means of students are valid alternatives. Furthermore, we could demonstrate that only 100 transcribed pages can already train an HTR system with reasonable performance. With the combination of naive transcribers, we believe that HTR systems can be quickly brought up to the performance of state-of-the-art systems trained with much larger datasets. A limitation of this study is still that it remains unclear how many naive transcribers need to be used and how many pages need to be transcribed. This is scope of future research. Furthermore, the current approach is based on a rather simple alignment and does not use any advanced methods that could potentially bring further error reduction.

References

Almazán, J., Gordo, A., Fornés, A., & Valveny, E. (2014). Word spotting and recognition with embedded attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36 (12), 2552–2566.

Bluche, T. (2016). Joint line segmentation and transcription for end-to-end handwritten paragraph recognition. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (NIPS) 29 (pp. 838–846). Curran Associates, Inc.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio,Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1724–1734). Association for Computational Linguistics.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In 1997 IEEE Workshop on Automatic Speech Rrecognition and Understanding (pp. 347–354).

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31 (5), 855–868.

Hillard, D., Hoffmeister, B., Ostendorf, M., Schlüter, R., & Ney, H. (2007). i ROVER: improving system combination with classification. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; companion volume, short papers (pp. 65–68).

Leifert, G., Strauß, T., Grüning, T., & Labahn, R. (2016). CITlab ARGUS for historical handwritten documents., arXiv:1605.08412

Maier, A., Hacker, C., Steidl, S., Nöth, E., & Niemann, H. (2005). Robust parallel speech recognition in multiple energy bands. In Joint Pattern Recognition Symposium (pp. 133–140).

Pitz, E. (1959). Schrift- und Aktenwesen der städtischen Verwaltung im Spätmittelalter. Mitteilungen aus dem Stadtarchiv von Köln, Bd, 45.

Rath, T. M., & Manmatha, R. (2003). Features for word spotting in historical manuscripts. In Seventh International Conference on Document Analysis and Recognition (ICDAR), 2003. Proceedings (pp. 218–222).

Rohlicek, J. R., Russell, W., Roukos, S., & Gish, H. (1989). Continuous hidden markov modeling for speaker-independent word spotting. In 1989 International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1989. icassp-89. (pp. 627–630).

Romero, E. G. (2017). Advances on the transcription of historical manuscripts based on multimodality, interactivity and crowdsourcing, (doctoral dissertation), doi: 10.4995/Thesis/10251/86137

Sánchez, J. A., Romero, V., Toselli, A. H., Villegas, M., & Vidal, E. (2017). Icdar2017 competition on handwritten text recognition on the read dataset. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 1383–1388).

Strauß, T., Weidemann, M., Michael, J., Leifert, G., Grüning, T., & Labahn, R. (2018). System description of CITLAB's recognition & retrieval engine for ICDAR2017 competition on information extraction in historical handwritten records; arXiv:1804.09943.

Sudholt, S., & Fink, G. A. (2016). PHOCNet: A deep convolutional neural network for word spotting in handwritten documents. In 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR) (pp. 277–282).