Sequence Learning in Unsupervised Training Cases

Martin Jenckel
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany.
Email: martin.jenckel@dfki.de

Syed Saqib Bukhari
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany.
Email: saqib.bukhari@dfki.de

Andreas Dengel
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany.
Email: andreas.dengel@dfki.de

I. Introduction

Sequence learning methods have been applied successfully to various recognition tasks like speech recognition, optical character recognition (OCR) or natural language processing. One of the most prominent methods is to use recurrent neural networks like Long-Short-Term-Memory (LSTM) networks, which have a memory unit that can store information through time. However most contemporary methods also require a large amount of annotated training data, which often is unavailable or very costly. Especially in the case of OCR on historical documents it requires many man-hours of language experts to transcribe and annotate the data for training.

A. Character Clustering

One way to apply supervised methods to unsupervised training cases is to generate the required annotated training data with the help of unsupervised methods like clustering. After segmenting the scanned documents into text lines and then single characters, clustering based on the differences in shape can be applied to the characters. The general challenges are choosing good features to represent the shapes and an algorithm that can cope with the inhomogeneous distribution of characters. For single characters of historical documents shown in Figure 1, clustering faces the additional challenge of degraded characters as shown in Figure 2. We have shown, that a combination of simple algorithm and features outperform more sophisticated approaches on this type of data [1].

B. anyOCR

The main focus of my research is the development of a self-learning end-to-end OCR system for historical documents called “anyOCR” [2]. The idea is to combine unsupervised methods, like clustering, with state-of-the-art sequence learning methods like LSTM-networks to provide a low effort and robust OCR system.
For clustering we use the idea of re-clustering bad results and came up with a fully automated algorithm, we call ”iterative k-means clustering”. The “anyOCR” approach still requires input from a language expert, but rather than annotating the full data, only the average image of each cluster has to be labeled. The results are then used as a erroneous ground truth for training the LSTM-network.
Due to the LSTMs inherent language modeling and context sensitive learning [3], it can correct many of the errors from the clustering. Testing on a 15^{th} century novel “Narrenschiff” in Latin script shown in Figure 1, we showed that our combination of methods can compete with approaches that use manually transcribed data. We could also show that relatively simple clustering can outperform more well regarded semi-supervised systems like Tesseract [4].

Fig. 1. Sample of the 15^{th} century novel “Narrenschiff” in Latin script from the German government funded project *Kallimachos*[2].

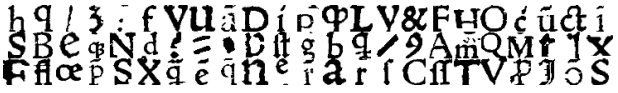

Fig. 2. Representatives of 69 character classes rescaled to 32x32 images. Some characters are incomplete or are otherwise degraded.

C. Future Work

In the future, besides further researching and improving the properties of LSTMs, we plan on extending the framework by combining it with other contemporary classification methods like convolutional neural networks (CNN) and look for applications in other classification domains.

References

[1] M. Jenckel, S. S. Bukhari, and A. Dengel, “Clustering benchmark for characters in historical documents,” in *DAS’16*, Greece, 2016.

[2] ——, “anyocr: A sequence learning based ocr system for unlabeled historical documents,” in *ICPR’16*, Mexico, 2016.

[3] T. M. Breuel, A. Ul-Hasan, M. Al Azawi, F. Shafait, “High Performance OCR for Printed English and Fraktur using LSTM Networks,” in *ICDAR*, Washington D.C. USA, aug 2013.

[4] A. UlHasan, S. S. Bukhari, and A. Dengel, “Ocroract: A sequence learning ocr system trained on isolated characters,” in *DAS’16*, Greece, 2016, pp. 174–179.

[2] http://kallimachos.de/kallimachos/index.php/Narragonien:Main