POS Tagging and Named Entity Recognition on Handwritten Documents

Vijay Rowtula, Praveen Krishnan, C.V. Jawahar
CVIT, IIIT Hyderabad
{vijay.rowtula, praveen.krishnan}@research.iiit.ac.in and jawahar@iiit.ac.in

Abstract

Parts of Speech (POS) tagging and Named Entity Recognition (NER) on handwritten document images can help in keyword detection during document image processing. In this paper, we propose an approach to detect POS and Named Entity tags directly from offline handwritten document images without explicit character/word recognition. We observed that POS tagging on handwritten text sequences increases the predictability of named entities and also brings a linguistic aspect to handwritten document analysis. As a pre-processing step, the document image is binarized and segmented into word images. The proposed approach comprising of a CNN-LSTM model, trained on word image sequences produces encouraging results on challenging IAM dataset.

1 Introduction

Information extraction from handwritten document images has numerous applications, especially in digitization of archived handwritten documents, assessing patient medical records and automated evaluation of student handwritten assessments, to mention a few. Document categorization and targeted information extraction from various such sources can help in designing better search and retrieval systems for handwritten document images. Keyword spotting (Fischer et al., 2012) is used for automatic document categorization by detecting the keywords or named entities directly on handwritten document images rather than transcribing to text to find keywords.

Semantic annotation of handwritten documents, especially spotting keywords using POS tags or NER is relatively a newer problem with very few works emerging on this front. In this paper, we attempt to fill the gap by proposing an approach for POS tagging and NER without handwriting transcription. The contribution of this work is to show generalization with a similar or improved performance of a unified end-to-end model without separating the sequence of sub-processes involved, thereby avoiding error propagation. Identifying named entities using noun phrases from POS tags can also be greatly helpful for keyword-based document retrieval. Detecting named entities irrespective of its structural and positional characteristics (eg. uppercase or lowercase letters) is an advantage of our approach. As a pre-processing step, we choose a handwritten dataset with segmented words and POS tag annotations. It helped us focus only on the aspect of POS and named entity tagging on handwritten word images rather than the problem of word segmentation from handwritten documents.

Related Works: Several state-of-the-art NER techniques were published in the literature using handcrafted features (Ritter et al., 2011; Lample et al., 2016). Transcription based models such as (Romero and Sánchez, 2013; Prasad et al., 2018; Carbonell et al., 2018) trained Handwritten Text Recognition (HTR) and NER jointly, to mitigate the disadvantage of errors in the first module affecting the next. But in historical handwritten documents, handwriting recognition struggles to produce an accurate transcription thereby reducing the accuracy of the whole system. Adak et al. (Adak et al., 2016) described an approach to directly detect the named entities from the document images. They used handcrafted features from document images with LSTM classifier, thereby avoiding the transcription step. The method relies on handcrafted features like identifying capital letters to detect possible named entities.

Figure 1: Example of POS and NE tagging on a sentence chosen from IAM handwritten dataset.

2 Our Approach

We hypothesize that, with sufficient handwritten document data and pre-processing, a deep learning model will be able to predict POS tags and named entities despite the inherent complexity, without the need for transcription.

2.1 Direct learning using synthetic dataset

Deep learning architectures need large datasets to attain decent results on image recognition tasks and finding sufficient handwritten document images is a challenging task. Hence we first trained the model on synthetic handwritten word images. We used a standard parts-of-speech dataset to create a synthetic handwritten dataset using artificial fonts, as described in (Krishnan and Jawahar, 2016). We used the same font for each sentence and sufficient data augmentation in the form of noise, translation, and rotation to resemble a large real handwritten dataset. Our assumption is that, with sufficient data, a deep learning model can generalize well on the end-to-end task without breaking it into sub-tasks (Liu et al., 2016). For POS tagging on handwritten text, our first step was to choose a model trained on word spotting in handwritten document images. The use of deep learning architectures to capture spatial features of word images is widely discussed in (Krishnan and Jawahar, 2016; Krishnan et al., 2016). The authors used HWNet architecture trained on 1 million word image dataset to make it robust to most handwriting variations. We initially used the pre-trained model (HWNet) to extract the features of synthetic handwritten words and, later fine-tuned a separate neural net on these features to classify POS tags. We considered this model was our baseline for the best performance that can be achieved using a pre-trained model on handwritten word images. In our alternate training scheme, we directly train a deep model on word images to classify POS tags. We observed that the model performance was similar to HWNet feature-based model which affirmed our assumption that translation into text or feature extraction sub-tasks may not be required for POS tagging on handwritten word images.

2.2 POS Tagging and NER

The model trained on the synthetic dataset is fine-tuned on a real handwritten dataset. We tested various architectures (CNN, CNN-LSTM) for both POS tagging and NER on a challenging handwritten document dataset. Some of them are discussed below.

Deep CNN model for POS tagging: Convolutional Neural Nets (CNN) are good in capturing the intricate details of images, hence making the model stable to inconsistencies like noise and translation (Krizhevsky et al., 2012). We trained a ResNet (He et al., 2016) model with 35 layers (validated empirically) on the synthetic dataset and fine-tuned it on IAM dataset for POS tagging task. The ResNet-35 ends with a softmax layer that outputs the probability distribution over the class la-

bels (POS tags). We trained the model with cross-entropy loss function to predict the class labels.

CNN-LSTM model for POS tagging: The probability of a transition between words may depend not only on the current observation, but also on past and future observations, if available (Lafferty et al., 2001). Since sentences in handwritten document images are word image sequences, we next used a combination of ResNet (CNN) and LSTM layers for training a POS tagging model on sequential information. We appended two layers of LSTM after ResNet-35 blocks and converted the input to LSTM as time distributed sequence. Different sequence lengths were tested on POS tags (classes). We report the performance of changing sequence lengths in the results section.

Named Entity Recognition: We adapt the similar architectures (CNN, CNN+LSTM) for the problem of NER. Here the underlying CNN architecture is ResNet-35. However, neither of the models had higher accuracy as noticed in similar experiments reported in (Toledo et al., 2016). We observed that named entities are related to position and distribution of POS tags in a sentence. We trained a multi-output classification network with architecture similar to POS model, with an extra branch of dense layers from the first fully connected dense layer, for named entity prediction. Hence the model now has an independent output with loss calculated from two sets of classes. As described in section 3.1, named entities have class imbalance problem. This is one of the reasons for choosing outputs separated by multiple dense layers rather than a common layer training for multi-class classification. We initially trained the network simultaneously for both POS and NER. We observed that though POS prediction accuracy remained the same as independent POS training, NER training did not give encouraging results. Hence we first trained the model (ResNet + LSTM + dense layers) for POS tagging by freezing the dense layers of NER. After the network achieved satisfactory accuracy on POS tagging, we froze the POS part of the network - including the ResNet-LSTM layers and trained just the dense layers of NER. We used altered class weights to tackle the class imbalance problem. This method improved the accuracy of NER better than any of the methods we have tried earlier.

Named Entities	**Tags**
Date	DATE
Geopolitical Entity	GPE
Organization	ORG
Person Name	PERSON
Nationalities or Religious or Political Groups	NORP
Unrelated	OTHERS
Not an Entity	–

Table 1: Named Entities used for our analysis.

3 Experimental Results and Discussions

Dataset: We used two different datasets, for training and fine-tuning the models. For training, a synthetic handwritten dataset was generated from chunking dataset of CoNLL-2000 shared task (Tjong Kim Sang and Buchholz, 2000), randomly using some of the 100 publicly available handwritten fonts (Krishnan and Jawahar, 2016). The chunking dataset contains sentences aligned with 211727 text tokens along with their POS tags in a separate train and test text files. This dataset was initially used for training and validation. The model is further fine-tuned on IAM handwritten dataset (Marti and Bunke, 2002). The IAM dataset contains 1539 forms written by 657 authors. The forms are further segmented into 115320 words and are annotated with POS tags. Though IAM dataset contains segmented lines and sentences, they are not properly annotated with text accordingly which makes it difficult to demarcate the individual sentences accurately. Hence we separated sentences based on pre-defined sentence rules based on words and cross-validated them using python based NLP tool named "Spacy".

Since the IAM handwritten forms have transcripts, the text was fed into the Spacy for generating the ground truth named entities. Spacy tagged sentences with 17 different categories of named entities. Though we restricted the classes to 6 named entities by choosing most recurrent tags, there was a class-imbalance problem. The list of tags used in this work is shown in Table 1. The unrelated entities occupied 92% of the NER classes. The IAM dataset is available as train, validation1, validation2, and test partitions. We used the training set to fine-tune our models and validated them against validation1 and validation2 sets.

Experiments	Precision	Recall	F1-score
Neural Net trained on HWNet features - CoNLL-2000 dataset synthetic images (POS tagging).	92.4	87.2	89.7
ResNet trained on - CoNLL-2000 dataset synthetic images (POS tagging).	94.2	84.5	89
ResNet trained on - CoNLL-2000 dataset synthetic images and fine-tuned on IAM dataset (POS tagging).	75.4	64.8	69.7
ResNet + LSTM trained on - CoNLL-2000 dataset synthetic images and fine-tuned on IAM dataset (POS tagging).	76.2	66.8	**71.2**
ResNet + LSTM trained on - CoNLL-2000 dataset synthetic images and fine-tuned on IAM dataset (NER).	74	64.1	**68.7**

Table 2: List of conducted experiments with precision, recall and F1-scores.

3.1 Results and Discussion

As a baseline on the synthetic dataset, we initially extracted HWNet features on word images from the fully connected layer and trained a multi-layered perceptron on 36 POS classes provided by CoNLL-2000 dataset. The model achieved an F1-score of 89.7. We then trained a 35 layer ResNet model which achieved an F1-score of 89. This was our initial experiment to prove that a model can be trained to classify POS tags directly on handwritten word images, rather than a feature based model training.

POS tagging on IAM dataset: The ResNet model trained and validated on the synthetic CoNLL-2000 dataset is fined tuned on IAM dataset. We initially trained directly on word images to classify 58 POS tags without the sequence information. The architecture essentially contained no LSTM layers. The ResNet model achieved an F1-score of 69.7 on IAM test dataset. We altered the architecture and dataset to include sequence information. We replaced dense layers succeeding the CNN layers with LSTM layers and trained the model with varying sequence lengths of 3, 64, 128 and 256 words. We observed that ResNet-LSTM model trained on 128 word length sequences performed best with an F1-score of 71.2 We attribute the decline of prediction accuracy on IAM dataset compared to synthetic dataset due to the following reasons. (i) Distortions in word images - We observed that most of the word images are formed by concatenating individual characters. (ii) Character distortions - characters such as '.' and ',' are displayed as 'l' in the dataset. (iii) Proper nouns errors - proper nouns do not start with capital letters. We also observed that 26% of errors were due to the noun form of words (NN), followed by adjectives (JJ) at 18% and conjunctions (IN, TO) at 12%. Rest of the errors were due to special characters, commas, and full stops.

NER on IAM dataset: Our models, training methods and metrics are summarized in table 2. We used class weights to bias the training towards named entity tags other than "unrelated" class, to handle class imbalance problem. We initially trained IAM dataset words for two tasks in parallel using the architecture described in section 2.2. But the accuracy of such model was low on NER task. Our first observation was that the errors caused by class imbalance were propagated back to the complete model which impacted the performance of both POS tagging and NER as well. Hence we first trained the model on POS tagging by freezing the NER layers, then we froze the layers for POS tagging and trained the model on NER. After 20 epochs, we fine-tuned the whole model further using very low learning rate for 10 epochs. The ResNet-LSTM model gave F1-score of 68.7 on NER on handwritten text.

4 Conclusion

A POS tagger and named entity recognizer for offline handwritten unstructured documents, without employing a character/word recognizer and an independent linguistic model, is presented in this paper. Experiments conducted on IAM dataset have resulted in an average F1-score of 71% on POS tagging and 68% on NER task. The proposed method is expected to work in other languages as well since our method deals with the linguistic aspect of handwritten documents where POS tags are identified first and then the NER.

References

Chandranath Adak, Bidyut B Chaudhuri, and Michael Blumenstein. 2016. Named entity recognition from unstructured handwritten document images. In *Document Analysis Systems (DAS), 2016 12th IAPR Workshop on*. IEEE.

Manuel Carbonell, Mauricio Villegas, Alicia Fornés, and Josep Lladós. 2018. Joint recognition of handwritten text and named entities with a neural end-to-end model. *arXiv preprint arXiv:1803.06252*.

Andreas Fischer, Andreas Keller, Volkmar Frinken, and Horst Bunke. 2012. Lexicon-free handwritten word spotting using character hmms. *Pattern Recognition Letters*, 33(7).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

Praveen Krishnan, Kartik Dutta, and CV Jawahar. 2016. Deep feature embedding for accurate recognition and retrieval of handwritten text. In *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on*. IEEE.

Praveen Krishnan and CV Jawahar. 2016. Matching handwritten document images. In *European Conference on Computer Vision*. Springer.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. 2016. End-to-end comparative attention networks for person re-identification. *arXiv preprint arXiv:1606.04404*.

U-V Marti and Horst Bunke. 2002. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1).

Animesh Prasad, Hervé Déjean, Jean-Luc Meunier, Max Weidemann, Johannes Michael, and Gundram Leifert. 2018. Bench-marking information extraction in semi-structured historical handwritten records. *arXiv preprint arXiv:1807.06270*.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Verónica Romero and Joan Andreu Sánchez. 2013. Category-based language models for handwriting recognition of marriage license books. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE.

Erik F Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*. Association for Computational Linguistics.

J Ignacio Toledo, Sebastian Sudholt, Alicia Fornés, Jordi Cucurull, Gernot A Fink, and Josep Lladós. 2016. Handwritten word image categorization with convolutional neural networks and spatial pyramid pooling. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer.