A Transcription Is All You Need: Learning to Align through Attention

Pau Torras, Mohamed Ali Souibgui, Jialuo Chen, and Alicia Fornés

Computer Vision Center, Computer Science Department
Universitat Autònoma de Barcelona
pau.torras@e-campus.uab.cat
{msouibgui, jchen, afornes}@cvc.uab.cat

Abstract. Historical ciphered manuscripts are a type of document where graphical symbols are used to encrypt their content instead of regular text. Nowadays, expert transcriptions can be found in libraries alongside the corresponding manuscript images. However, those transcriptions are not aligned, so these are barely usable for training deep learning-based recognition methods. To solve this issue, we propose a method to align each symbol in the transcript of an image with its visual representation by using an attention-based Sequence to Sequence (Seq2Seq) model. The core idea is that, by learning to recognise symbols sequence within a cipher line image, the model also identifies their position implicitly through an attention mechanism. Thus, the resulting symbol segmentation can be later used for training algorithms. The experimental evaluation shows that this method is promising, especially taking into account the small size of the cipher dataset.

Keywords: Handwritten symbol alignment · Hand-drawn symbol recognition · Sequence to Sequence · Attention Models

1 Introduction

Historical ciphered manuscripts have recently attracted the attention of many researchers [6], not only for their own historical value, but also because of the challenges related to the transcription, decryption and interpretation of their contents. Indeed, many of these ciphers apply different techniques to hide their content from plain sight, for example, by using invented symbol alphabets. An example of a ciphered manuscript [1] is illustrated in Fig. 1.

Transcribing the sequence of symbols in the manuscript is the first step in the decryption pipeline [9]. However, machine learning-based recognition methods require annotated data, which is barely available. Indeed, an accurate labelling (e.g. annotation at symbol level) is desired, since it can then be used for training symbol classification, segmentation, spotting methods, etc. But, the few expert transcriptions are often available at paragraph or line level. For this

[1] https://cl.lingfil.uu.se/~bea/copiale/

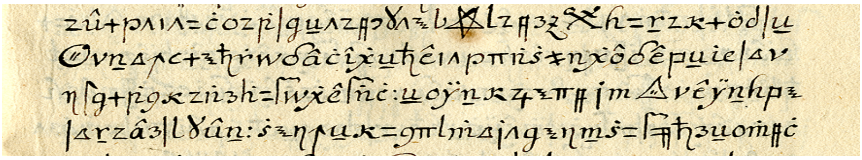


Fig. 1. An example of the Copiale ciphered manuscript, related to an 18th-century German secret society, namely the "oculist order".

reason, we propose to align each transcribed symbol with its representation in the manuscript image by using an attention-based Seq2Seq model [4], which implicitly infers the position of relevant visual features for every character output step.

The rest of the paper is organized as follows. First, in Section 2 we delve into relevant alignment methods present in the literature. We describe our approach in detail in Section 3, and the experiments in Section 4. Finally, in Section 5 we present some future work avenues and a few closing words.

2 Related Work

Many approaches exist for the task of alignment, which vary depending on the nature of the aligned manuscript. An example of domain-specific alignment can be found in Riba *et al.* [7], which consists of an image-to-image alignment using Dynamic Time Warping for detecting variations in music score compositions without the need of transcriptions. Similarly, Kassis *et al.* [5] use Siamese Neural Networks to align two handwritten text images with the same contents but different writing style.

The image-to-text alignment has been also researched. For example, Romero *et al.*[8] use Hidden Markov Models (HMM) and a dynamic programming algorithm to find candidate transcriptions of text lines and align them to the ground truth sequence. Fischer *et al.* [3] use HMMs and a first recognition pass.

A combination of both approaches is proposed by Ezra *et al.* [2], where they overfit a recogniser on the input data and generate a synthetic version of the image, which is aligned. They also use the OCR output and edit distances between said output and the ground truth for better performance.

3 Proposed Method

In this section we describe our architecture. We have used the Seq2Seq model with an attention mechanism proposed by Kang *et al.*[4] for HTR (Handwritten Text Recognition). We have adapted this technique for the task of alignment and performed several modifications related to the way attention masks are presented to improve their accuracy and flexibility for aligning cipher symbols.

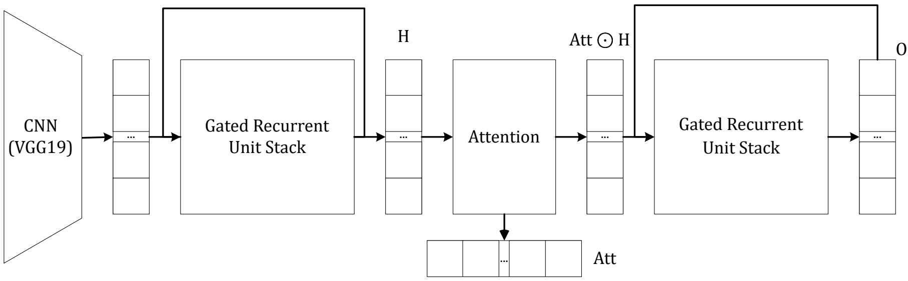


Fig. 2. Representation of the Seq2Seq model and the placement of attention within the pipeline.

3.1 Sequence to Sequence model

The Seq2Seq model is an Encoder-Decoder architecture, which means that it processes an input set of vectors sequentially, generates an intermediate representation from them and then generates an output sequence based on said representation. The addition of an attention mechanism makes it possible for the intermediate representation to contain an unset number of vectors, since the model learns to assert the relevance of each of them and conditioning on those most useful in the current step. Our Seq2Seq model, depicted in Fig. 2, has a VGG19 convolutional network with its last max pooling layer removed as the first step in the pipeline, which accepts a $800 \times 64 \times 3$ px image of a text line as input. This generates a $50 \times 4 \times 512$-element (flattened to 50×2048) representation, which the Encoder, a stack of Gated Recurrent Units (GRU), further annotates into the hidden state H. Then, for each Decoder inference step, a vector Att of dimension 50 is computed through an attention mechanism. The input for the Decoder, another GRU stack, is the Hadamard product between the Att vector and the hidden state H, concatenated with the previously inferred symbol.

Our hypothesis is that there is a direct correlation between the position of a highly active attention mask and the position of the associated inferred symbol in the output text sequence, which enables us to perform alignment when learning to recognise lines.

3.2 Attention Mask Tuning

When applying the original Seq2Seq model in our data we found a major limitation, which was the fact that the attention mechanism can only provide a discrete set of fixed positions of a set width, since the attention mask is a 50-element vector that represents relevant areas in an 800px wide image. This made segmenting narrow characters or long sequences very difficult. Moreover, since the output of the attention mechanism is the result of a Softmax layer, no more than one attention band per character has a significant value.

Thus, we improved the model by treating the attention mask as a histogram and fitting a Normal Distribution onto it in order to find the position and width of the character in a more precise manner. Thus, every character mask is computed as:

$$m_{low} = \mu - kc_l\sigma \quad m_{high} = \mu + kc_h\sigma + 1 \tag{1}$$

where m_{low}, m_{high} are the lower and higher bounds of the character mask, μ is the mean of the histogram, σ is the standard deviation, c_l and c_h are a distribution skewness correction factor and k is a scaling factor to account only a set fraction of the standard deviation.

In this work, the k value was set to 0.5 after a tuning process in order to have only the most relevant samples of the distribution within the set boundaries. The skewness correction factors were computed as the ratio between the sum of all bins before or after the mean divided by the total sum of bins (excluding the highest valued one). Note also that both m_{low} and m_{high} need to be converted from the attention mask coordinate space into the image coordinate space.

Finally, since the model's only input is an image, the quality of the alignment relies on the model's capacity to produce a good output sequence of tokens. And given that our goal is not recognising the image, but instead aligning the associated transcription, the final mask prediction can be corrected by finding the shortest edit path between the output and ground truth sequences using Levenshtein's algorithm to remove unnecessary masks or adding padding when required. This underlying assumption considering the shortest edit path corresponds to the sequence of mistakes that a model has actually made, is quite strong, but we found that it prevents (or, at the very least, alleviates) misalignment in the majority of cases.

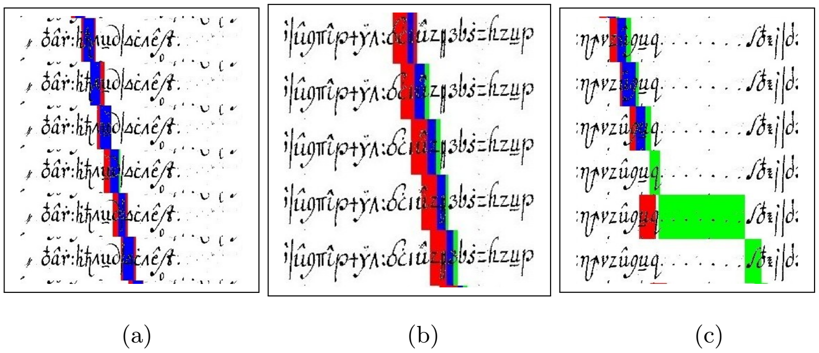
4 Experiments and Results

In this section we present the experiments performed to assess the viability of our model. We trained the Seq2Seq recogniser using line-level samples from the Copiale cipher with early stopping at 30 epochs with no Symbol Error Rate (SER) improvement on validation. Table 1 includes relevant information about the dataset and the model's hyperparameters.

For comparison, we used as baseline a learning-free method [1], which segments the line into isolated symbols based on connected components analysis. First, connected components are extracted, and then, grouping rules are join the components that likely belong to the same symbol. Then, the components are aligned to the sequence of transcribed symbols.

Table 1. Relevant training information for experiment reproducibility.

Optimiser	Adam	**Training Samples**	649
Learning Rate (LR)	$3 \cdot 10^{-4}$	**Validation Samples**	126
LR Checkpoints	@ 20, 40, 60, 80, 100 epochs	**Test Samples**	139
LR Sigma	0.5	**Dataset Classes**	126
Loss Function	Cross-Entropy	**Avg. Line Length**	42



(a) (b) (c)

Fig. 3. Qualitative Results. Model's prediction in red, ground truth in green and the intersection of both in blue color. Each successive line within the image is a time step. These are fragments of a longer alignment sequence, cut for readability purposes. (a), (b) and (c) are examples of output quality patterns.

Quantitative results are shown on Table 2. Segmentation accuracy is evaluated as the Intersection over Union (IoU): the percentage of masks whose ground truth and prediction intersect in a ratio of t over the union of both areas. As it can be seen, our approach surpasses the baseline method in most scenarios.

The analysis of these results shows three general patterns:

- **Correct alignment** (Fig. 3a) : There is a considerable proportion of cases with an overall correct alignment with limited error. Perfect masks are however rare, with some degree of error on the sides being relatively frequent.
- **Slight misalignment** (Fig. 3b) : Mostly caused due to incorrect edit paths chosen after the recognition algorithm, narrow symbols or very long sequences, which cause the attention masks to be broader in comparison.
- **Misalignment** (Fig. 3c): Incorrect alignment when encountering very rare symbols or high output SER, mostly due to limited training data.

Finally, we note that we tested the potential for bootstrapping training under the same parameters with synthetic samples. We created a 40.000 sample dataset using segmented symbols from real Copiale pages. We trained the model with random-width lines with each symbol appearing under a uniform distribution

Table 2. Experimental results, considering different Intersection over Union (IoU) thresholds. Metrics are Precision (Prec.), Recall (Rec.), F_1 score and Average Precision (AP). Symbol Error rate of the classifier is 0.365.

	IoU t=0.25				**IoU t**=0.50				**IoU t**=0.75			
Exp.	**Prec.**	**Rec.**	F_1	**AP**	**Prec.**	**Rec.**	F_1	**AP**	**Prec.**	**Rec.**	F_1	**AP**
Baseline	0.31	0.30	0.31	0.12	0.25	0.24	0.24	0.07	0.14	0.16	0.14	0.02
Ours	0.94	0.89	0.91	0.84	0.59	0.55	0.57	0.34	0.10	0.09	0.10	0.001

and fine-tuned it real samples. However, results did not improve, which we attribute to characters in synth lines being broader than their real counterparts, which caused attention masks to skip symbols.

5 Conclusion

We have proposed an alignment method based on Seq2Seq models. Our method shows encouraging results given the small dataset. The main hindrances are the difficulty for training a very accurate model and the need of further mask processing in order to be able to find bounding boxes correctly. Thus we believe that, by refining the Levenshtein algorithm including confidence data to choose the right edits or modifying the attention mechanism to have more than one high activation mask, results might improve. It is also worth to explore adding supervised attention mask training to avoid having to tune masks after recognition, since some character-level annotated samples are available for the data we are working with and we might boost the performance further.

Acknowledgement

This work has been supported by the Swedish Research Council, grant 2018-06074, DECRYPT – Decryption of Historical Manuscripts, the Spanish project RTI2018-095645-B-C21 and the CERCA Program / Generalitat de Catalunya.

References

1. Baró, A., Chen, J., Fornés, A., Megyesi, B.: Towards a generic unsupervised method for transcription of encoded manuscripts. In: DATeCH. pp. 73–78 (2019)
2. Ezra, D.S.B., Brown-DeVost, B., Dershowitz, N., Pechorin, A., Kiessling, B.: Transcription alignment for highly fragmentary historical manuscripts: The dead sea scrolls. In: ICFHR. pp. 361–366 (2020)
3. Fischer, A., Frinken, V., Fornés, A., Bunke, H.: Transcription alignment of latin manuscripts using hidden markov models. In: HIP. pp. 29–36 (2011)
4. Kang, L., Toledo, J.I., Riba, P., Villegas, M., Fornés, A., Rusinol, M.: Convolve, attend and spell: An attention-based sequence-to-sequence model for handwritten word recognition. In: German Conference on Pattern Recognition. pp. 459–472 (2018)
5. Kassis, M., Nassour, J., El-Sana, J.: Alignment of historical handwritten manuscripts using siamese neural network. In: ICDAR. vol. 1, pp. 293–298 (2017)
6. Megyesi, B., Esslinger, B., Fornés, A., Kopal, N., Láng, B., Lasry, G., de Leeuw, K., Pettersson, E., Wacker, A., Waldispühl, M.: Decryption of historical manuscripts: the decrypt project. Cryptologia **44**(6), 545–559 (2020)
7. Riba, P., Fornés, A., Lladós, J.: Towards the alignment of handwritten music scores. In: GREC. pp. 103–116 (2015)
8. Romero-Gómez, V., Toselli, A.H., Bosch, V., Sánchez, J.A., Vidal, E.: Automatic alignment of handwritten images and transcripts for training handwritten text recognition systems. In: DAS. pp. 328–333 (2018)
9. Souibgui, M.A., Fornés, A., Kessentini, Y., Tudor, C.: A few-shot learning approach for historical ciphered manuscript recognition. In: ICPR. pp. 5413–5420 (2021)