A comparative study of information extraction strategies using an attention-based neural network

Solène Tarride, Aurélie Lemaitre, Bertrand B. Coüasnon, Sophie Tardivel

A comparative study of information extraction strategies using an attention-based neural network

Solène Tarride[1,2], Aurélie Lemaitre[1], Bertrand Coüasnon[1], and Sophie Tardivel[2]

[1] Univ. Rennes, IRISA, CNRS, France
[2] Doptim, Cesson-Sévigné, France

Abstract. This article focuses on information extraction in historical handwritten marriage records. Traditional approaches rely on a *sequential* pipeline of two consecutive tasks: handwriting recognition is applied before named entity recognition. More recently, *joint* approaches that handle both tasks at the same time have been investigated, yielding state-of-the-art results. However, as these approaches have been used in different experimental conditions, they have not been fairly compared yet. In this work, we conduct a comparative study of sequential and joint approaches based on the same attention-based architecture, in order to quantify the gain that can be attributed to the joint learning strategy. We also investigate three new joint learning configurations based on multi-task or multi-scale learning. Our study shows that relying on a joint learning strategy can lead to an 8% increase of the complete recognition score. We also highlight the interest of multi-task learning and demonstrate the benefit of attention-based networks for information extraction. Our work achieves state-of-the-art performance in the ICDAR 2017 Information Extraction competition on the Esposalles database at line-level, without any language modelling or post-processing.

Keywords: Document image analysis · Historical documents · Information extraction · Handwriting recognition · Named entity recognition

1 Introduction

In recent years, many European libraries, museums and archives have undertaken to digitize their collections of historical documents [20], as a way to ensure the preservation of our cultural heritage. Indeed, digital documents do not need to be physically handled, thus reducing the potential damage to fragile collections, and can be quickly and easily accessed from any location. However, searching for a specific document among millions of digitized entries remains a challenge. As a result, there is a need for data indexation, which would allow users to browse a collection and retrieve specific documents using keyword queries. A practical application is the search of specific information contained in population records. These documents hold a strong value for genealogists around the world, as they

provide precious information about our ancestors. The ability to search for a specific name in a collection of records would substantially ease genealogical research. Additionally, a macro-analysis of such documents could provide an interesting perspective for historians and demographers.

This raises the question of how to extract information from large collections of historical records. Several platforms have been proposed for collaborative manual annotation, such as Transcribathon[3] and CrowdHeritage[4]. Yet, even with collaborative tools, manual extraction of information from large collections still requires a lot of effort. Recently, the advances in computer vision have opened the way for automatic document understanding using computational models. These systems can be trained to recognize a sequence of handwritten characters, and to assign a semantic label to each predicted word. However, these approaches are not entirely reliable when applied to challenging documents, such as historical documents. Indeed, historical records can have various layouts and handwriting styles, and often feature paper and ink degradations, abbreviations, inter-line annotations, or crossed-out words.

The interest for the task of information extraction in historical documents has been reinforced by the 2017 Information Extraction in Historical Handwritten Records (IEHHR) competition [12], based on Esposalles database [24]. The database is a collection of historical handwritten marriage records from the Archives of the Cathedral of Barcelona from the 17th century. The aim of the competition is to extract relevant information about the wife and husband and their parents, such as their name, occupation and place of origin, as illustrated in Fig. 1.

Fig. 1. Example from the database proposed for the IEHHR competition [12]. Each image is associated with a transcription, and two semantic labels are associated with each word: *category* and *person*.

Most of the approaches submitted to this competition are based on a sequential approach, in which the handwriting recognition task is performed upstream of the named entity recognition task [12]. However, several researchers have highlighted the possibility of combining these two tasks, by training a model to output

[3] https://www.transcribathon.com/en/
[4] https://crowdheritage.eu/en

characters and contextual tags [7, 25]. These joint approaches yield competitive results, however it is impossible to assert whether the improvement comes from the architecture used by the authors or from the joint learning strategy.

In this work, we conduct a comparative study of these two training strategies using the same attention-based model, in order to quantify the improvement linked to the joint learning strategy. We also introduce three additional joint learning configurations based on multi-task and multi-scale learning.

This article is organized as follows. Section 2 introduces the works related to automatic information extraction and handwriting recognition in historical documents. Section 3 introduces the attention-based sequence-to-sequence architecture proposed in this work. Section 4 presents our comparative study and introduces three joint learning strategies based on multi-task and multi-scale learning. The results of our experiments are presented and discussed in Section 5. Finally, we summarize this work and propose future directions.

2 Related Works

Automatic document understanding is a challenging research area that includes document layout analysis, handwriting recognition, and information extraction. As we work on pre-segmented line images, we focus this study on handwritten text recognition (HTR) and information extraction (IE).

2.1 Handwriting recognition

State-of-the art HTR models are currently based on deep neural networks. One of the most popular architectures is composed of a convolutional neural network (CNN) to extract features from the image, and a recurrent neural network (RNN) to capture sequential information. This architecture is often referred to as CRNN-CTC, as it is trained using the Connectionist Temporal Classification (CTC) loss function [13]. This architecture has become very popular over the recent years [5, 14, 23]. More recently, attention-based sequence-to-sequence (seq2seq) networks have been investigated for HTR. Contrary to the CRNN-CTC architecture, attention-based models learn to align image pixels with the target sequence. As a result, the network learns to focus on a small relevant part of the feature vector to predict each token [19, 22]. Another strength of this architecture is that the recurrent decoder learns an implicit language model at character-level. The seq2seq architecture has also demonstrated its ability to handle multiple tasks at once [16], which is interesting in the context of information extraction. Finally, the Transformer architecture is also gaining a lot of attention from the HTR community [15], although this architecture requires many training images.

2.2 Information extraction

The task of information extraction (IE) consists in extracting semantic information from documents. For structured documents, such as tables, forms or invoice,

semantic information can be derived from the word localization. In this case, it is common to build end-to-end models that localize each word, transcribe them, and derive context from localization features [21, 30]. But for semi-structured documents, such as marriage records, context can only be derived from phrasing. For example, a surname generally comes after a name. In this scenario, the first step is to localize textual zones, such as paragraphs, text-lines or words. Then, handwriting recognition and named entity recognition (NER) is performed.

Three main strategies have been considered so far. The traditional strategy is a *sequential approach* where handwriting recognition is performed before named entity recognition [12]. Another *sequential approach* consists in classifying each word into semantic categories, then applying handwriting recognition techniques [27]. The last strategy is a *joint approach*, where both tasks are tackled at the same time, using an end-to-end model [6, 7].

Transcription before word semantic classification The most common approach relies on a HTR model to predict a transcription, and on natural language processing (NLP) techniques to classify each word into named entities using textual features. The drawback of this sequential approach is that there is no contextual information during the transcription stage. Rather, the context is used in the post-processing stage, using category-based language modelling. Several methods have been proposed in the ICDAR2017 competition on Information Extraction [12]. The Hitsz-ICRC team developed an approach at word-level. First, a bi-gram based CNN is trained on word images to recognize characters, without any language model. Then, words are classified using the CRF sequence tagging method. The CITlab ARGUS team developed an approach at line-level. Images are passed into a CRNN-CTC architecture for handwritten text recognition, then regular expressions are used to decode and classify each word.

Word semantic classification before transcription This approach consists in labelling each word before predicting the transcription. The interest is that knowing the semantic category beforehand helps the system to predict the right transcription. It is particularly suited when dealing with word image, as each image can be easily classified before being transcribed. Toledo et al. [27] extract semantic context from word images, and obtain the transcription of each word using semantic context of precedent words. The authors observe that their system benefits from knowing the semantic category. For example, if it makes sense to read a male name, the word "John" will be more likely than the word "born", even if the handwritten word looks more like "born". The main drawback of this approach is that it requires word bounding box annotation.

Joint transcription and word semantic classification This approach consists in producing a transcription and a semantic category at the same time. A method has been proposed in [6], where an end-to-end model is used to localize word bounding boxes and jointly classifying and transcribing them. A major drawback is that it requires word-level segmentation for training. Another approach

based on line-level images [7] relies on a CRNN-CTC network to predict a transcription with semantic categories, using tags located before important words. If this approach is very promising, we believe that it could benefit from using an attention-based network, which would allow the network to learn the appropriate features to predict the tags. More recently, this joint approach has been successively used with a Transformer network at line-level and record-level [25]. We believe that this combined approach is very promising, and benefits from an attention-based network to contextual features that are relevant to predict the semantic tags.

2.3 Our statement

In this work, we address different aspects of information extraction at line-level.

First, we want to assess the interest of the joint approach for information extraction. Our intuition is that combining handwriting recognition and named entity recognition should be helpful, as both tasks are related and could benefit from shared contextual features. Indeed, semantic context can be derived from the transcription, but knowing the semantic context beforehand should also make the transcription more reliable. Recently, competitive results have been achieved using this joint approach [7, 25], which indicates its relevance for information extraction. However, it is challenging to assert if the improvement comes from the neural network architecture or from the combination of HTR and NER. To validate this intuition, we compare sequential and joint approaches using the same neural network. Our study aims to measure the improvement that can be attributed to the learning strategy.

Secondly, we present a simple seq2seq architecture with attention, and demonstrate the interest of attention-based models for the task of information extraction. Indeed, the attention mechanism learns to focus on relevant zones of the image at each step, which should facilitate the recognition of named entities based on visual features. Moreover, this architecture can learn an implicit language model, which is convenient for semi-structured documents. To this end, we compare its performance with other neural network architectures, mainly CRNN-CTC [7] and Transformer [25].

Finally, we investigate three additional joint learning strategies based on our seq2seq architecture using multi-task and multi-scale learning. Indeed, we believe that these strategies can be helpful for joint handwriting recognition and named entity recognition, as they have been proved efficient for neural translation and image captioning [16].

Our contributions are summarized as follows:

- We show the benefit of joint strategies for information extraction and handwriting recognition.
- We highlight the interest of attention-based models for joint handwriting and named entity recognition.
- We propose three additional joint strategies for HTR and NER using multi-task and multi-scale learning strategies.

– We obtain state-of-the-art results on the IEHHR competition benchmark at line-level [12], without any post-processing.

3 The attention-based seq2seq architecture

The seq2seq architecture is an encoder-decoder network which was initially proposed for automatic translation [26]. It recently gained popularity for speech recognition, image captioning and neural translation [3, 8, 29]. The architecture has since been adapted for HTR [19, 22], as illustrated in Fig. 2. In this work, we use an architecture adapted from [19].

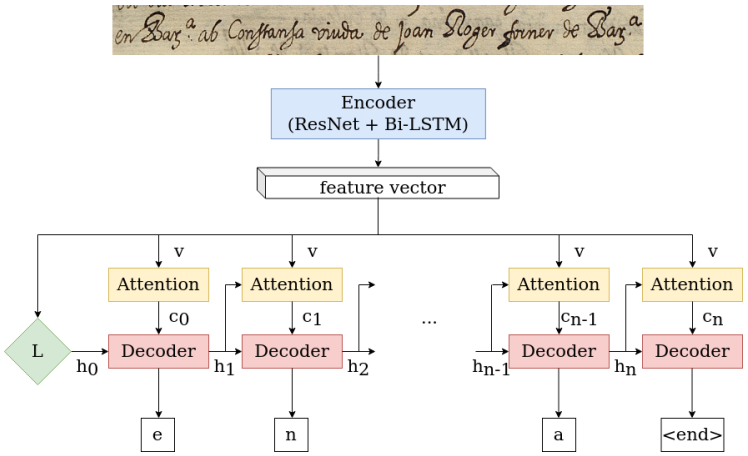


Fig. 2. Seq2seq architecture used in this work, where v is the feature vector, L is a linear function, (c_i) and (h_i) are the context vector and the decoder's hidden state at time step i, and n is the number of decoding steps.

Input Line images are pre-processed and augmented before entering the encoder. First, they are resized, padded and normalized. Then, random augmentations are applied on the fly, including DPI adjusting, random perspective, random transform, elastic distortion, dilation, erosion, sign flipping, brightness and contrast adjustment.

Encoder The role of the encoder is to extract features from the image. Our encoder is a convolutional recurrent neural network (CRNN) based on a ResNet-101 backbone, which is pre-trained on ImageNet [10]. Only the last two convolutional blocks are trained to reduce the number of training parameters. We add to this architecture a bi-dimensional Long Short Time Memory (BLSTM) network to capture the sequential information. The final feature vector is then fed to the attention mechanism and the decoder.

Attention The attention mechanism has been introduced by Bahdanau et al. [3] as a way to focus on parts of the feature vector, without any manual segmentation. For every step, the attention mechanism allows the network to rely on contextual features that are useful for the task at hand. Several attention mechanisms have been proposed over the years [3,8,17]. In this work, we use Chorowski attention (hybrid) [8], as it yields better results for HTR [19].

Decoder The role of the decoder is to generate a textual sequence from the attention-aware feature vector. The decoder is a simple LSTM cell trained with an embedding layer. For each time step, the output is passed through a softmax layer to obtain probabilities for each character of the alphabet.

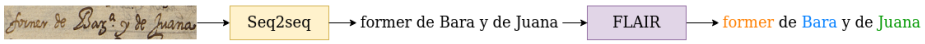
Training settings The seq2seq is trained using the hybrid loss proposed in [19], and using the Adam optimization algorithm. We use teacher forcing in the decoder during training. Early stopping is used to stop the training if no improvement is observed on the validation loss for 20 epochs. During inference, we use the best validation weights and beam search decoding with a beam size of 5.

4 Strategies for information extraction

In this section, we present the two main approaches compared in this work. We also introduce three additional learning strategies for joint handwriting and named entity recognition. No language modelling or post-processing is used in this work.

4.1 Comparing the *sequential* and *joint* approaches

We present the two main strategies for information extraction: the *sequential approach* and the *joint approach*. These two approaches are illustrated in Fig. 3 and will be compared using the same seq2seq architecture in Section 5.



(a) *Sequential approach*: the seq2seq network is used to predict characters and FLAIR is used to classify predicted words into named entities.

(b) *Joint approach*: the seq2seq network is used to predict characters and tags

Fig. 3. Illustration of the two main strategies compared in this work. Legend: wife's father occupation, wife's father location, wife's mother name. Figure best viewed in color.

Sequential approach This strategy is the most popular approach for information extraction [12]. Handwriting recognition and named entity recognition are addressed as separate subsequent tasks, as illustrated in Fig. 3a. Once handwritten text recognition is performed, each predicted word is classified into multiple semantic categories, based on textual features. Our implementation of this approach relies on our seq2seq network for handwriting recognition, and the FLAIR system [1] for named entity recognition, trained using catalan word embedding and FLAIR embedding.

Joint approach This approach, illustrated in Fig. 3b, was initially proposed in [7] using a CRNN-CTC neural network. In this strategy, the network is trained to predict tags as well as characters, with tags being located before each word of interest. In this work, we propose to reproduce this experiment using our seq2seq architecture trained to predict characters and tags. We believe that using an attention-based architecture is appropriate, as the network learns to focus on relevant parts of the feature vector to predict the tags. The tags are designed to encode the semantic category and the person relative to each word. We use the combined tags proposed in [7], e.g. <wife_name> is used to represent the wife's name.

4.2 Exploring additional *joint* learning configurations

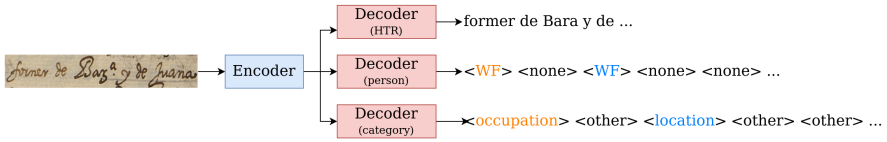
Finally, we propose and evaluate three original *joint* configurations using multi-task and multi-scale training. Our intuition is that learning multiple tasks, or from multiple scales, could help the encoder to extract richer contextual features.

Joint multi-task strategy without tags In this first learning configuration, three decoders are connected to the same encoder, as illustrated in Fig. 4a. Each decoder is trained for a specific task:

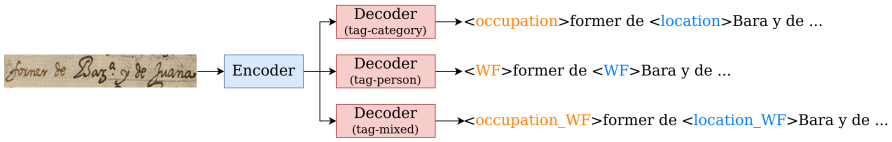
- the *htr* decoder predicts the sequence of characters
- the *category* decoder predicts the sequence of categories (6 classes: name, surname, location, occupation, state, other)
- the *person* decoder predicts the sequence of persons (8 classes: wife, father, wife's father, wife's mother, husband's father, husband's mother, other person, none)

Each decoder relies on a specific attention mechanism that focuses on relevant features for each subtask. The network is trained using a single loss function that is computed as the mean of the three individual loss functions. During inference, the sequences are merged to assign a category and a person to each predicted word.

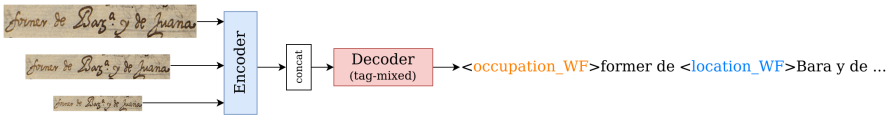
However, this approach has two possible issues. The first is a potential convergence issue, as the *htr* decoder typically learns slower than the other decoders. The second issue comes from the alignment of the predicted sequences: an error at the beginning of a sequence can potentially offset the entire prediction.



(a) *Joint multi-task without tags*: A single encoder is connected to three decoders. Each decoder is specialised for the prediction of characters, categories and persons.



(b) *Joint multi-task with tags*: A single encoder is connected to three decoders. Each decoder is specialised for the prediction of category-based tags, person-based tags, and mixed tags.



(c) *Joint multi-scale strategy*: the image is fed to the encoder at multiple scales.

Fig. 4. Illustration of three *joint* strategies proposed in this work. Legend: WF for wife's father

Joint multi-task strategy with tags This second learning configuration is designed to overcome the issues of the last strategy. It is illustrated in 4b. Three decoders are connected to the same encoder, and each decoder is trained using a specific semantic encoding:

- *tag-category* is a single tag for the category, e.g <name>
- *tag-person* is a single tag for the person, e.g <wife>
- *tag-mixed* is a combined tag for both category and person, e.g <wife_name>

Each decoder relies on a specific attention mechanism that focuses on relevant features for each subtask. The network is trained using a single loss function that is computed as the mean of the three individual loss functions. During inference, only the *tag-mixed* branch is evaluated to ensure a fair comparison with the single-task *joint* strategy. In this scenario, the subtasks tackled by each decoder are balanced, which ensure proper convergence.

Joint multi-scale strategy with tags The multi-scale learning configuration is based on the *joint* strategy and is illustrated in Fig. 4c. In this scenario, the image is passed through the encoder at different scales to get contextual information at different levels. The three feature vectors are concatenated and passed through the attention network and the decoder.

5 Experiments

In this section, we evaluate our contributions for information extraction. First, we present results achieved by our attention-based network for handwriting recognition, showing that it is competitive. Then, we conduct a comparative study of *sequential* and *joint* approaches for information extraction based on our attention-based network, and evaluate three additional joint learning configurations. Finally, we compare this work with other participants in the IEHHR competition.

5.1 Handwriting recognition using seq2seq

First, we demonstrate the interest of our seq2seq architecture for HTR, as it is used for the *sequential strategy*. The evaluation is carried out on two public databases at line-level: IAM [18] (modern, English) and RIMES [2] (modern, French). We use the Character Error Rate (CER) as the evaluation metric. We compare our architecture with other state-of-the-art methods, without any post-processing or language model, in Tab. 1. Our architecture yields state-of-the-art results on the IAM database. When comparing our seq2seq (encoder-decoder) with the corresponding CRNN-CTC (encoder-CTC) architecture, we observe that the seq2seq architecture is more efficient. Indeed, when using the attention-based model, the CER drops from 6.1% to 5.2% on the IAM database, and from 5.2% to 4.4% on the RIMES database. This is likely due to attention mechanism and the implicit language model learned by the decoder. Consequently, we use the seq2seq architecture in the following experiments.

Table 1. Benchmark comparison of methods for handwriting recognition at line-level on the test set and without language modeling or post-processing.

System	Method	IAM	RIMES
CRNN-CTC	Wigington et al. [28]	6.4	**2.1**
CRNN-CTC	Puigcerver [23]	5.8	2.3
CRNN-CTC	Dutta et al. [11]	5.2	5.1
CRNN-CTC	*Ours*	6.1	5.2
Seq2seq	Poulos et al. [22]	16.6	12.1
Seq2seq	Chowdhury et al. [9]	8.1	3.5
Seq2seq	Bluche [4]	7.9	2.9
Seq2seq	Michael et al. [19]	**5.2**	-
Seq2seq	*Ours*	**5.2**	4.4
Transformers	Kang et al. [15]	7.6	-

5.2 Information extraction using seq2seq

We analyse the four approaches described in Section 4 for information extraction on the Esposalles database. Then, we compare our best approaches with state-of-the-art methods.

Evaluation protocol Two tasks can be evaluated using the Esposalles database [24]. The first task is full text recognition, which is evaluated using the Character Error Rate (CER) and the Word Error Rate (WER). The second task is information extraction, which can be assessed using the metrics introduced in the ICDAR 2017 competition on Information Extraction in Historical Handwritten Records (IEHHR) [12]. The objective of information extraction is to go beyond full text recognition, by assigning two semantic labels to each word. For the basic track, the aim is to recognize the category among *name*, *surname*, *occupation*, *location*, *civil state* and *other*. The score associated to the basic task is $100 - CER$ if the category is correctly identified, 0 otherwise. For the complete track, the role of the person must also be identified among: *wife*, *husband*, *wife's father*, *wife's mother*, *husband's father*, *husband's mother*, *other person*, *none*. The score associated to the complete task is $100 - CER$ if both labels are correctly identified, 0 otherwise. It should be noted that only relevant words (e.g. not classified as *other*/*none*) are taken into account in this evaluation.

Comparing the sequential and joint approaches The results presented in Tab. 2 compare the two main strategies presented in this work. The results highlight the interest of the *joint* strategy for information extraction task, as it yields a 3% increase on the basic score and an 8% increase on the complete score of the IEHHR competition.
We observe that the *sequential* approach tends to propagate and amplify transcription errors. Indeed, we have observed that words with errors are more likely to be misclassified: the complete score drops from 98.5% when FLAIR is applied to ground truth words to 86.7% when applied to predicted words. Another weakness of the *sequential* approach can be observed on the complete score. The score drops by 5% when the evaluation of persons is considered, suggesting that this task is particularly problematic for the FLAIR model at line level.
Overall, the strength of the *joint* approach comes from the implicit knowledge regarding the context of each word during the prediction. We observe that this knowledge helps the handwriting recognition task, as the CER drops from 2.82% when predicting only characters to 1.81% when predicting characters and tags.

Evaluation of additional joint learning configurations We observe that the *multi-task without tags* approach obtains a high error rate, which confirms our intuition on the asymmetric convergence of the three branches: the two *ner* decoders converge much quicker than the HTR decoder. This causes either an overfitting of the two *ner* decoders or an underfitting of the *htr* decoder. Moreover, difficulties related to the alignment of the three predicted sequences may explain the low scores on the IEHHR competition. An interesting contribution of our study is the

Table 2. Comparison of various learning strategies for information extraction on the Esposalles database at line-level. The first table presents results for handwriting recognition and information extraction using scores from the IEHHR competition. The second table details these scores for each semantic category.

	CER	WER	Basic score	Complete score
Sequential	2.82	8.33	91.2	86.7
Joint	1.81	6.10	94.7	94.0
Joint multi-task without tags	7.75	17.38	61.8	48.1
Joint multi-task with tags	**1.74**	**5.38**	**95.2**	**94.4**
Joint multi-scale with tags	5.61	15.13	83.0	80.3

	Name	Surname	Location	Occupation	State
Sequential	94.6	85.9	91.7	92.3	90.4
Joint	96.1	91.0	93.7	**95.3**	**97.8**
Joint multi-task without tags	63.2	41.0	48.2	69.8	86.7
Joint multi-task with tags	**97.0**	**92.6**	**94.5**	**95.3**	96.7
Joint multi-scale with tags	87.4	61.1	84.2	87.5	96.9

joint multi-task with tags strategy. This learning strategy yields better results than the *joint* strategy, even though we only evaluate the *tag-mixed* branch. This highlights that the network benefits from learning different semantic representations at once. This result also confirms the observation of Luong et al. [16] who observed that multi-task learning improves the performance of seq-to-seq models for neural translation. However, we must relativize the small improvement of this strategy with its high computational cost. Finally, the *joint multi-scale with tags* strategy does not meet our expectations. Our intuition was that extracting features at different scales would help to get more contextual information. However, we observe that this makes it more difficult for the attention network to select relevant features. As a result, the final performance is quite poor.

Benchmarking information extraction (IEHHR) We now compare our work with other participants in the IEHHR competition in Tab. 3. Our *joint multi-task with tags* strategy achieves state-of-the-art results at line-level, without any post-processing and language model. This result shows the interest of multi-task learning for information extraction, despite a high computational cost.
Another interesting point comes from the comparison of our *joint* strategy with the articles that rely on the same strategy using a different architecture, mainly CRNN-CTC [7] and Transformer [25]. When compared to the CRNN-CTC [7], our seq2seq model is able to boost the complete score from 89.40% to 94.0% using the same methodology. When compared to the Transformer [25], our seq2seq model obtain competitive results, with a lower basic score but a higher complete score. This observation highlights the interest of attention-based architectures, such as seq2seq or Transformer, for information extraction.

Table 3. Benchmark comparison of the IEHHR competition on the Esposalles database at line level.

Method	Model	Strategy	Basic score	Complete score
Baseline HMM [12]	HMM+LM	Sequential	80.2	63.1
CITlab-ARGUS-1 [12]	CRNN-CTC	Sequential	89.5	89.2
CITlab-ARGUS-2 [12]	CRNN-CTC	Sequential	91.9	91.6
CITlab-ARGUS-3 [12]	CRNN-CTC	Sequential	91.6	91.2
CVC (tags) [7]	CRNN-CTC	Joint	90.6	89.4
InstaDeep (tags) [25]	Transformer	Joint	**95.2**	93.3
Joint (ours)	Seq2seq	Joint	94.7	94.0
Joint multi-task with tags (ours)	Seq2seq	Joint	**95.2**	**94.4**

Another valuable observation comes from the analysis of the attention maps. We observe that the attention is very narrow and well focused on the corresponding pixels when predicting characters. However, when the network predicts semantic tags, the attention spreads over the previous words, showing that it learns to attend over relevant visual features to predict semantic categories.
It should be noted that a work submitted on the IEHHR competition website[5] outperforms our approaches, although the methodology has not been published yet. Moreover, the joint approach based on a Transformer model [25] achieves state-of-the-art results at record-level, as the network benefits from a larger bi-dimensional context.

6 Conclusion

This study compares joint and sequential approaches for information extraction. Our results demonstrate the interest of using *joint* approaches, as we show that training a network for handwriting recognition and named entity recognition increases performance on both tasks. Compared to the traditional *sequential* approach, our *joint* strategy yields to an 8% increase in complete recognition score and a significant decrease of the Character Error Rate (from 2.82% to 1.81%). In addition, this work highlights the interest of seq2seq architectures. Indeed, we obtain a substantial performance increase when the *joint* strategy is applied using a seq2seq network, as compared to the CRNN-CTC approach [7]. This is because seq2seq networks rely on an attention mechanism to extract relevant visual features, as well as a recurrent decoder to learn the implicit language model. Its performance is comparable to the Transformer proposed in [25], although our approach yield a better complete score. Finally, we explore different joint learning configurations and observe that multi-task learning from multiple semantic encodings helps the network to extract relevant features for each task. Indeed, our *joint multi-task with tags* approach yields a complete score of 94.4% on the IEHHR competition at line-level. As a consequence, we believe

[5] https://rrc.cvc.uab.es/?ch=10

that multi-task seq2seq architectures should be investigated in more depth. We obtain state-of-the-art results on the IEEHR competition [12], without any post-processing or external language modeling. We believe that future work should also focus on information extraction at paragraph-level to take advantage of the recurrent phrasing at record-level. Recent work shows that relying on incremental learning strategies could ease information extraction at record-level [25].

Acknowledgements

Solène Tarride is partly funded by the CIFRE ANRT grant No. 2018/0896.

References

1. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: NAACL Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). pp. 54–59 (2019)
2. Augustin, E., Brodin, J.m., Carré, M., Geoffrois, E., Grosicki, E., Prêteux, F.: RIMES evaluation campaign for handwritten mail processing. In: Proc. of the Workshop on Frontiers in Handwriting Recognition (2006)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate (2016)
4. Bluche, T.: Joint Line Segmentation and Transcription for End-to-End Handwritten Paragraph Recognition (2016)
5. Bluche, T., Messina, R.: Gated Convolutional Recurrent Neural Networks for Multilingual Handwriting Recognition. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 646–651 (2017)
6. Carbonell, M., Fornés, A., Villegas, M., Lladós, J.: A Neural Model for Text Localization, Transcription and Named Entity Recognition in Full Pages. Pattern Recognition Letters **136** (2020)
7. Carbonell, M., Villegas, M., Fornés, A., Lladós, J.: Joint Recognition of Handwritten Text and Named Entities with a Neural End-to-end Model. CoRR **abs/1803.06252** (2018)
8. Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-Based Models for Speech Recognition. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. p. 577–585. NIPS'15, MIT Press, Cambridge, MA, USA (2015)
9. Chowdhury, A., Vig, L.: An Efficient End-to-End Neural Model for Handwritten Text Recognition. CoRR **abs/1807.07965** (2018)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
11. Dutta, K., Krishnan, P., Mathew, M., Jawahar, C.: Improving CNN-RNN Hybrid Networks for Handwriting Recognition. In: 16th International Conference on Frontiers in Handwriting Recognition. pp. 80–85 (2018)
12. Fornés, A., Romero, V., Baró, A., Toledo, J.I., Sánchez, J.A., Vidal, E., Lladós, J.: ICDAR2017 Competition on Information Extraction in Historical Handwritten Records. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) **01**, 1389–1394 (2017)

13. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In: Proceedings of the 23rd International Conference on Machine Learning. p. 369–376. ICML '06, Association for Computing Machinery, USA (2006)
14. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A Novel Connectionist System for Unconstrained Handwriting Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(5), 855–868 (2009)
15. Kang, L., Riba, P., Rusiñol, M., Fornés, A., Villegas, M.: Pay Attention to What You Read: Non-recurrent Handwritten Text-Line Recognition. CoRR **abs/2005.13044** (2020)
16. Luong, M.T., Le, Q., Sutskever, I., Vinyals, O., Kaiser, L.: Multi-task Sequence to Sequence Learning. Proceedings of ICLR, San Juan, Puerto Rico (11 2015)
17. Luong, M.T., Pham, H., Manning, C.D.: Effective Approaches to Attention-based Neural Machine Translation (2015)
18. Marti, U.V., Bunke, H.: The IAM-database: An English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition **5**, 39–46 (11 2002)
19. Michael, J., Labahn, R., Grüning, T., Zöllner, J.: Evaluating Sequence to Sequence Models for Handwritten Text Recognition. CoRR **abs/1903.07377** (2019)
20. Nauta, G.J., Heuveland, W.V.D., Teunisse, S.: Europeana DSI 2–Access to Digital Resources of European Heritage - Report on ENUMERATE Core Survey 4 (2017)
21. Palm, R.B., Laws, F., Winther, O.: Attend, Copy, Parse End-to-end Information Extraction from Documents. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 329–336 (2019)
22. Poulos, J., Valle, R.: Character-Based Handwritten Text Transcription with Attention Networks. CoRR **abs/1712.04046** (2017)
23. Puigcerver, J.: Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition? In: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 01, pp. 67–72 (2017)
24. Romero, V., Fornés, A., Serrano, N., Sánchez, J.A., Toselli, A., Frinken, V., Vidal, E., Lladós, J.: The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognit. **46**, 1658–1669 (2013)
25. Rouhou, A.C., Dhiaf, M., Kessentini, Y., Salem, S.B.: Transformer-based approach for joint handwriting and named entity recognition in historical document. Pattern Recognition Letters (2021)
26. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to Sequence Learning with Neural Networks. In: Proc. NIPS. Montreal, CA (2014)
27. Toledo, J.I., Carbonell, M., Fornés, A., Lladós, J.: Information extraction from historical handwritten document images with a context-aware neural model. Pattern Recognition **86**, 27–36 (2019)
28. Wigington, C., Tensmeyer, C., Davis, B.L., Barrett, W., Price, B.L., Cohen, S.: Start, Follow, Read: End-to-End Full-Page Handwriting Recognition. In: ECCV (2018)
29. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. CoRR **abs/1502.03044** (2015)
30. Yu, W., Lu, N., Qi, X., Gong, P., Xiao, R.: PICK: Processing Key Information Extraction from Documents using Improved Graph Learning-Convolutional Networks. 2020 25th ICPR pp. 4363–4370 (2021)