

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

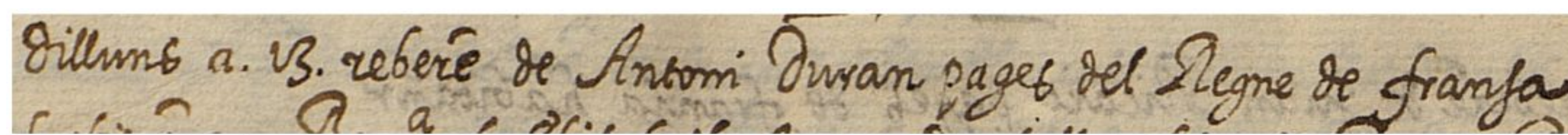
1. INTRODUCTION

1.1 Context

- We focus on handwritten historical records
- Demographic records have **high historical value** for genealogists and historians
- These documents have been **digitized, but cannot be searched** using keyword queries
- We aim to develop an automatic strategy to **extract important information** from these documents

1.2 Database

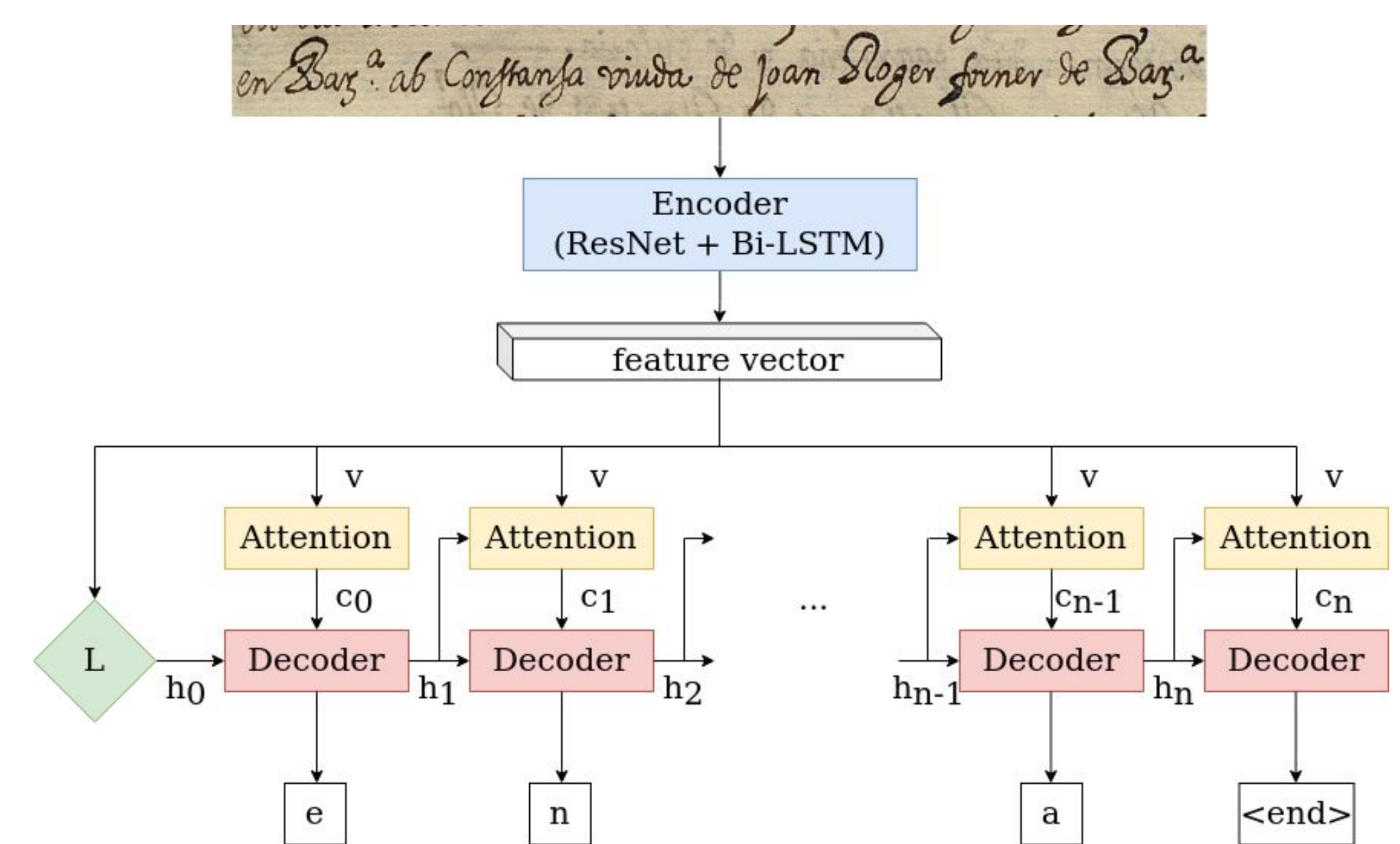
- We work on the Esposalles database [1]
 - historical documents** from the Archives of the Cathedral of Barcelona from the 17th century
 - marriage records written by **one writer** in old Catalan
- Aim of the IEHHR competition [2]
 - handwritten text recognition (HTR)
 - named entity recognition (NER) with **semantic categories** and **persons**
- 100 pages available for training and validation, 25 pages for testing



dilluns	a	13	reberere	de	Antoni	Duran	pages	del	Regne	de	fransa
other	other	other	other	other	name	surname	occupation	other	location	location	location
none	none	none	none	none	husband	husband	husband	none	husband	husband	husband

2. OUR ATTENTION-BASED MODEL FOR HTR AND IE

- Traditional models are based on a **CRNN-CTC** architecture
- We propose to use an attention-based **seq2seq** architecture
 - extracts **relevant visual features** using the attention mechanism
 - learns an **implicit language model** in the decoder
- State-of-the-art** results on IAM and Esposalles at line level without any post-processing or language model



3. EXPLORING STRATEGIES FOR INFORMATION EXTRACTION

3.1 Contributions

- We compare fairly **sequential and joint learning strategies** for information extraction
- We introduce an original strategy based on a **multi-task seq2seq** network inspired by [3].
- We use the **same seq2seq architecture** with attention mechanism **without any post-processing or language model**

3.2 Results

Approach	Basic score (%) ↑	Complete score (%) ↑	CER (%) ↓	WER (%) ↓
Sequential	91.2	86.7	2.82	8.33
Joint	94.7	94.0	1.81	6.10
Joint multi-task	95.2	94.4	1.74	5.38

4. CONCLUSION

4.1 Summary of our contributions

- Sequence-to-sequence neural networks with an **attention mechanisms are adapted for information extraction** in historical records
- Joint HTR and NER** using contextual tags **improve recognition**, as compared to traditional sequential approaches
- Multi-task strategies are well suited for this task** as multiple specialized decoders share contextual information through the encoder
- Our system achieves **state-of-the-art performance** on the IEHHR competition **at line level**.

Approach	Basic score (%) ↑	Complete score (%) ↑
CITlab-ARGUS-2 [2]	91.9	91.6
CVC [4]	90.6	89.4
InstaDeep [5]	95.2	93.3
Joint (ours)	94.7	94.0
Joint multi-task (ours)	95.2	94.4

4.2 Future works

- Working at **paragraph-level** would allow the network to get **more contextual information**
- Exploring other neural networks with attention (Transformer)

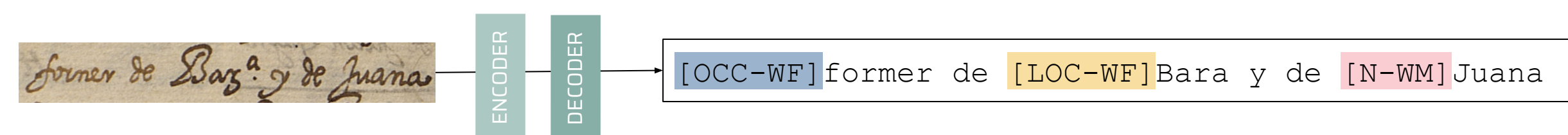
Sequential approach

- Method 1: HTR+FLAIR** - Seq2seq for HTR then FLAIR for NER



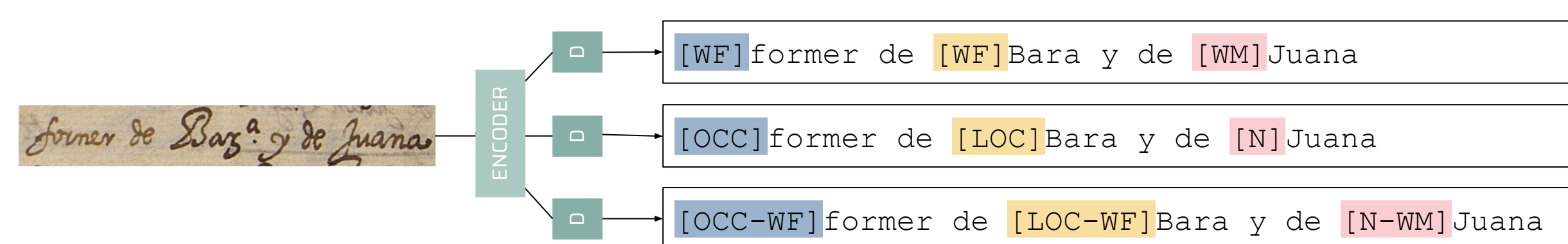
Joint approach

- Method 2: Tags** - Seq2seq for characters and tags



Joint multi-task approach

- Method 3: Tags multi-task** - Multi-task seq2seq for characters and tags



Legend: Wife's father occupation - Wife's father location - Wife's mother name | N = name; OCC = occupation; LOC = location; WF = wife's father; WM = wife's mother

Solène TARRIDE^{1,2},
Aurélie LEMAITRE²,
Bertrand COÜASNON²,
Sophie TARDIVEL¹

¹Doptim, ²Univ Rennes (France)

Contact: solene.tarride@irisa.fr

References

- Verónica Romero, Alicia Fornés, Nicolás Serrano, Joan Andreu Sánchez, Alejandro H. Toselli, Volkmar Frinken, Enrique Vidal, & Josep Lladós (2013). The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognition, 46(6), 1658-1669.
- Alicia Fornés, Verónica Romero, Arnau Baró, Juan Ignacio Toledo, Joan Andreu Sánchez, Enrique Vidal, & Josep Lladós (2017). ICDAR2017 Competition on Information Extraction in Historical Handwritten Records. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (pp. 1389-1394).
- Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, & Lukasz Kaiser. (2016). Multi-task Sequence to Sequence Learning.
- Manuel Carbonell, Mauricio Villegas, Alicia Fornés & Josep Lladós (2018). Joint Recognition of Handwritten Text and Named Entities with a Neural End-to-end Model.
- Ahmed Cheikh Rouhoua, Marwa Dhiab, Yousri Kessentini, & Sinda Ben Salem (2021). Transformer-Based Approach for Joint Handwriting and Named Entity Recognition in Historical documents. CoRR, abs/2112.04189.