Evaluation of Named Entity Recognition in handwritten documents

David Villanova-Aparisi[1], **Carlos-D. Martínez-Hinarejos**[1], **Verónica Romero**[2], **Moisés Pastor-Gadea**[1]
[1] PRHLT Research Center, Universitat Politècnica de València, Camí de Vera, s/n, València 46021, Spain
[2] Departament d'Informàtica, Universitat de València, València 46010, Spain

Main contributions

- Proposal of two evaluation metrics for the combined Handwritten Text Recognition (HTR) and Named Entity Recognition (NER) task
- Application of syntactical constraints to improve the performance of a coupled model

Previous work

- Coupled approaches avoid error propagation in several tasks
- Usage of Convolutional-Recurrent Neural Networks (CRNN)
- Previous experimentation on the chosen dataset

Evaluation metrics

Classic Measures

- CER and WER:
  - Tags are considered as characters or words
  - Impossible to evaluate the syntactical correctness
- Precision, Recall and F1-Score:
  - Specific focus on Named Entities
  - Problems with multiple appearances and order constraints
  - Cannot adjust the strictness of the metric

Entity CER and Entity WER

Edit distance with operation costs:

$$\mathrm{I}(i,j) = 1$$
$$\mathrm{D}(i,j) = 1$$
$$\mathrm{S}_{\mathrm{CER}}(i,j) = \begin{cases} 2 & \text{if } E_i \neq E_j \\ 2 \cdot \mathrm{CER}(T_i, T_j) & \text{otherwise} \end{cases}$$
$$\mathrm{S}_{\mathrm{WER}}(i,j) = \begin{cases} 2 & \text{if } E_i \neq E_j \\ 2 \cdot \mathrm{WER}(T_i, T_j) & \text{otherwise} \end{cases}$$

Main benefits:

- Specific focus on Named Entities
- Consideration of order constraints
- Deals with multiple appearances
- The strictness can be adjusted

Experimental method

Dataset:

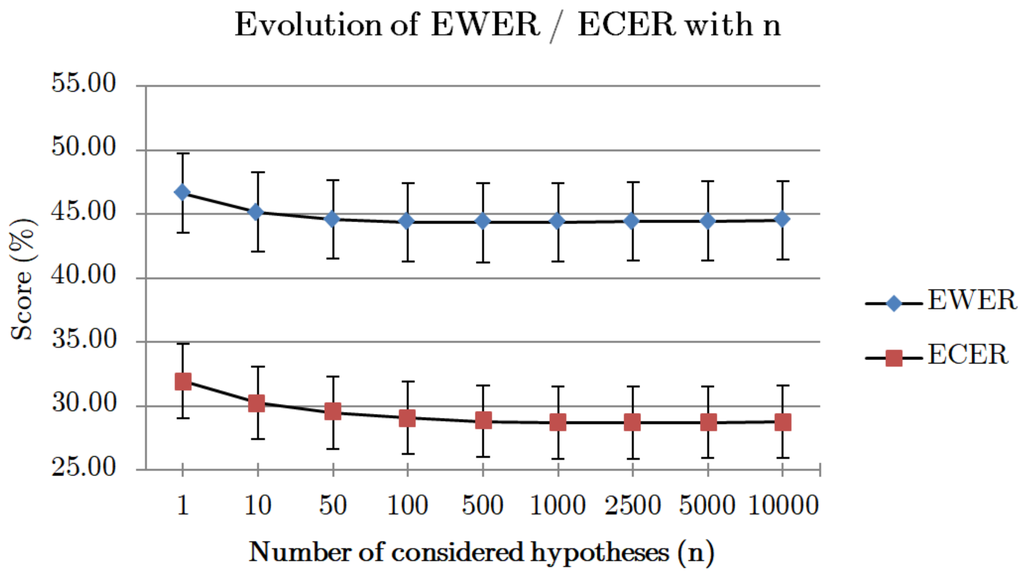
- 499 letters written by different authors
- Three languages: Latin, Czech and German
- Types of Named Entities: Person, Place and Date
- Parenthesized notation and nested Named Entities
- Data partitioning:
  - Training set: 398 letters (80%)
  - Validation set: 51 letters (10%)
  - Test set: 50 letters (10%)

Employed architecture:

- Optical Model: CRNN implemented and trained with PyLaia
- Language Model: Character 8-gram estimated with SRILM
- Combination of both models via Kaldi
- Decoding: Obtain the first syntactically correct hypothesis among the n-best outputs

Obtained results

Metric	Boroş, Emanuela et al. (no nested NEs)	Combined model (1-best, nested NEs)	Combined model (2500-best, nested NEs)
CER (%)	**8.00 ±1.68**	9.23 ±1.80	9.24 ±1.80
WER (%)	**26.80 ±2.75**	28.20 ±2.79	28.14 ±2.79
Precision (%)	**49.25 ±3.10**	43.14 ±3.07	40.05 ±3.04
Recall (%)	37.08 ±3.00	37.58 ±3.00	**39.97 ±3.04**
F1 (%)	**42.30 ±3.07**	40.17 ±3.04	40.01 ±3.04
ECER (%)	34.48 ±2.95	31.94 ±2.89	**28.69 ±2.81**
EWER (%)	52.79 ±3.10	46.62 ±3.10	**44.42 ±3.08**



Conclusions

- Two novel metrics for the combined task based on edit distance
- Increase of the number of syntactically correct outputs
- No statistically significant improvements over our baseline system

Future work

- Consideration of Named Entites spanning over several lines
- Paragraph level decoding
- Apply our approach in different corpora

Acknowledgements

This work was supported by Grant RTI2018-095645-B-C22 funded by MCIN/AEI/ 10.13039/501100011033, by "ERDF A way of making Europe", by Grant ACIF/2021/436 funded by Generalitat Valenciana, by Generalitat Valenciana under the project GV/2021/072 and by Generalitat Valenciana under project DeepPattern (PROMETEO/2019/121).