La Rochelle Université

ASSESSING AND MINIMIZING THE IMPACT OF OCR QUALITY ON NAMED ENTITY RECOGNITION

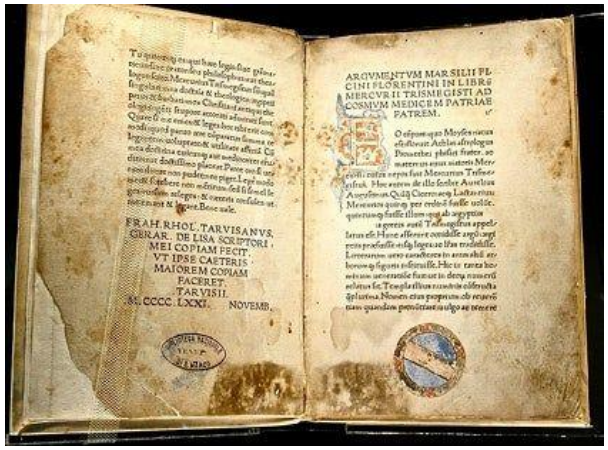
Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidère, Mickaël Coustaty and Antoine Doucet

TPDL 2020

NEWS E E

Motivation

- In digital libraries, documents are digitized and archived as images.
- The accessibility to their textual content requires an OCR processing.
- OCR errors due to the quality of documents, storage conditions...
- Named entities are the first point of entry for users in a search system.
- 4/5 user queries on the Gallica digital library contain at least one named entity.



The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Named Entity Recognition

Named Entity Recognition (NER) is the task that aims to locate named entities in a text and to categorize them into a set of predefined classes.

A **Named Entity (NE)** is a real-world objects that refers to a unique entity.

Classes of NEs: person PER, location LOC, organization ORG, human product PROD, miscellaneous MISC

Examples

- *Paris Hilton stayed at the Hilton in Paris.*
- *The New York Times is an American newspaper based in New York City.*

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NER approaches

La Rochelle Université

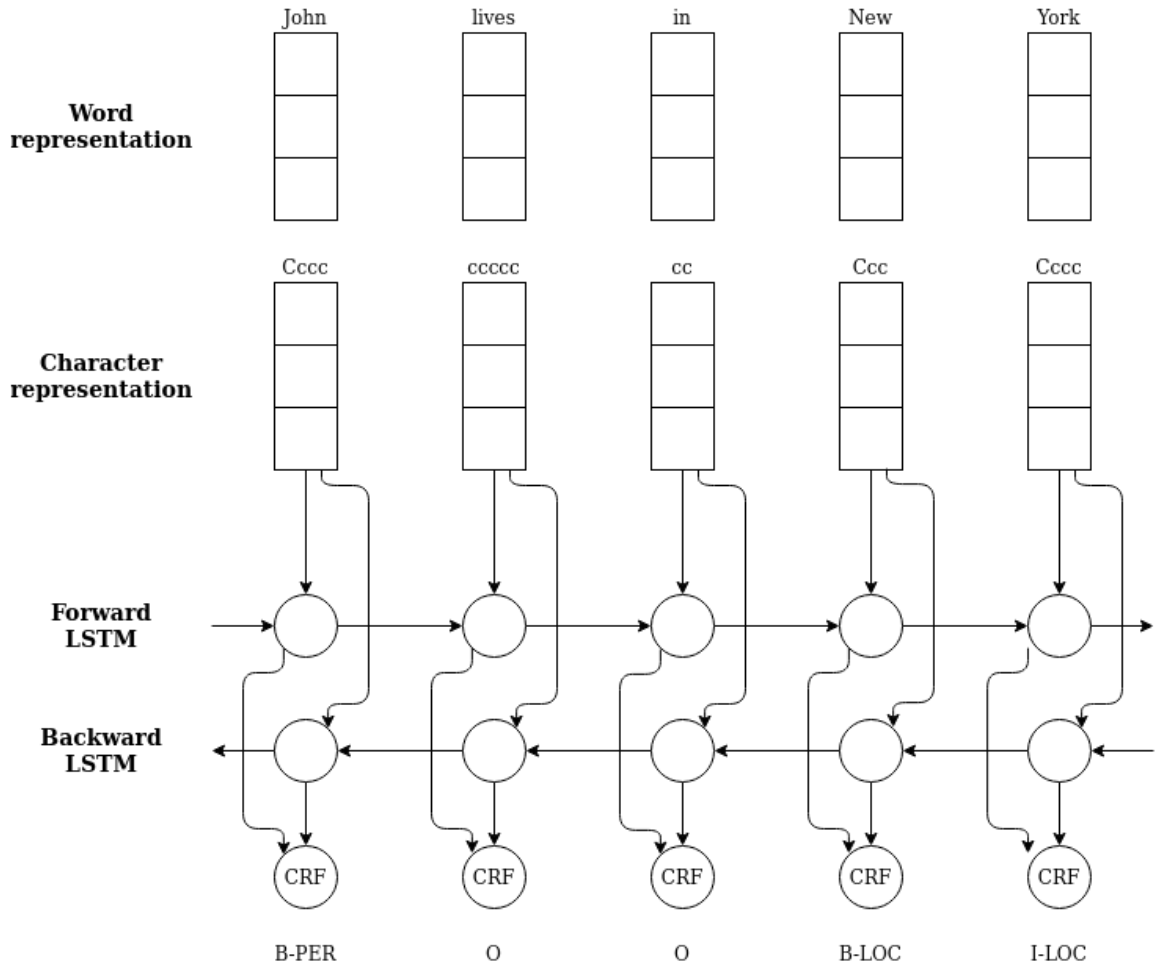
• **Rule-based approach**: rules (mainly defined manually) are related to lexica of proper names, linguistic descriptions and trigger words.

• **Machine learning-based approach**: extract rules automatically based on learning systems trained on large corpora. Since 2011, neural networks showed an ability to outperform the previous NER system.

1. CoreNLP (Stanford NER)
2. BLSTM-CRF
3. BLSTM-CNN
4. BLSTM-CNN-CRF

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

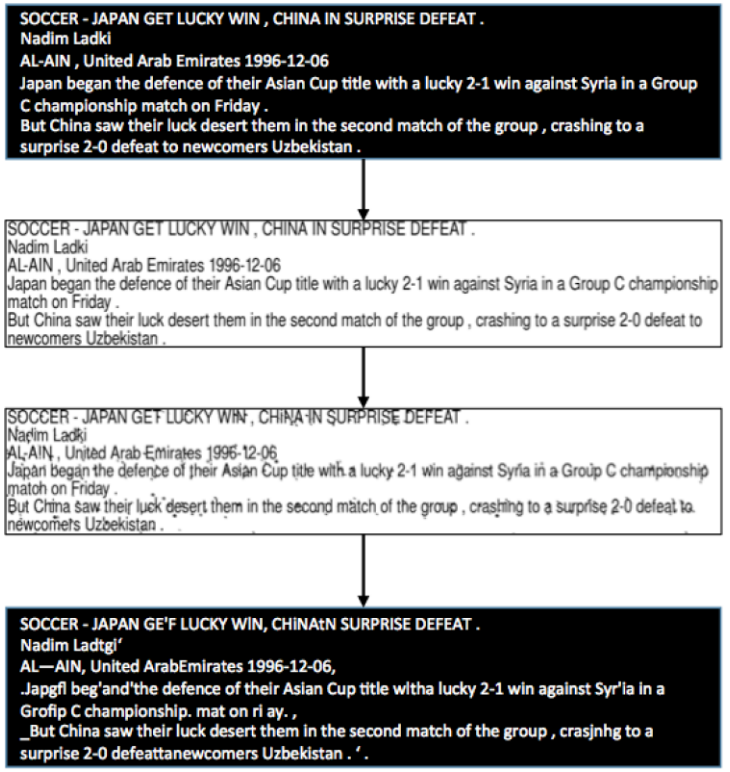
BLSTM models



The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Synthetic document degradation

- No noisy NER corpora aligned with their clean versions
- Available clean NER data sets:
  1. CONLL-2002: Spanish and Dutch
  2. CONLL-2003: English
- Injection of OCR degradation



Four degradations

- Character degradation
- Phantom degradation
- Blurring
- Bleeding effect

Two levels

- Rare
- Reasonably frequent

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Text alignment

Alignment of degraded and original texts by tool RETAS:

OCR : SOCCER - JAPAN GE'F@ LUCKY Wl@N@, CHi@NAt@@N SURPRISE DEFEAT . Nadim Ladtg@i 'AL—@AIN@, United Arab@
GT : SOCCER - JAPAN GE@@T LUCKY W@IN , CH@INA@ IN SURPRISE DEFEAT . Nadim Lad@@ki @AL@-AIN , United Arab

OCR : Emirates 1996-12-06, . Japgfl@@ beg'and'@the defence of their Asian Cup title wl@th@a lucky 2-1 win a
GT : Emirates 1996-12-06@ @@Jap@@an beg@an@@ the defence of their Asian Cup title w@ith a lucky 2-1 win a

OCR error rates:

- **Character error rate (CER)**: substitution, deletion and insertion
- **Word error rate (WER)**: an erroneous word has at least one erroneous character

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

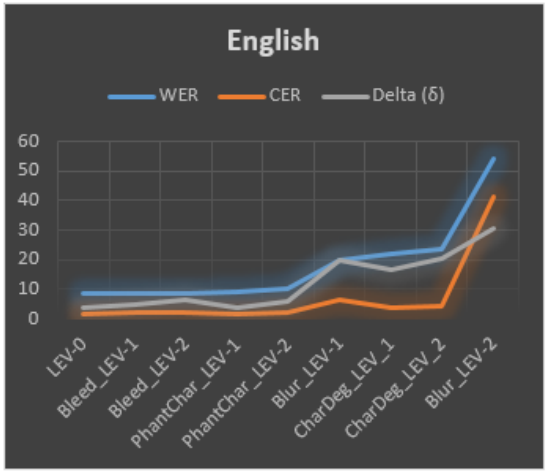
OCR error rates

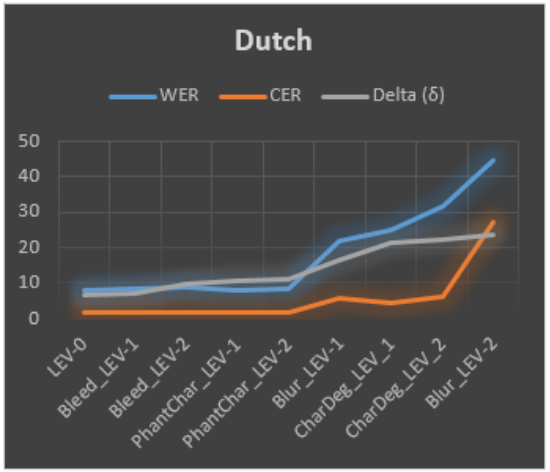
		English		Dutch		Spanish	
		CER	WER	CER	WER	CER	WER
LEV-0		1.7	8.5	1.6	7.8	0.7	4.8
Bleed-through	LEV-1	1.8	8.5	1.7	8.2	0.8	4.9
	LEV-2	1.8	8.6	1.8	8.9	0.8	5.4
Blurring	LEV-1	6.3	20.0	5.9	22.0	3.0	12.0
	LEV-2	41.3	54.0	27.0	44.7	19.5	29.9
Char deg.	LEV-1	3.6	21.8	4.5	25.1	2.1	14.2
	LEV-2	4.3	23.7	6.4	31.6	2.7	16.3
Phantom deg.	LEV-1	1.7	8.8	1.6	8.0	0.8	5.5
	LEV-2	1.8	10.0	1.7	8.4	0.9	5.9
LEV-MIX		6.9	22.8	5.8	22.2	3.5	11.9

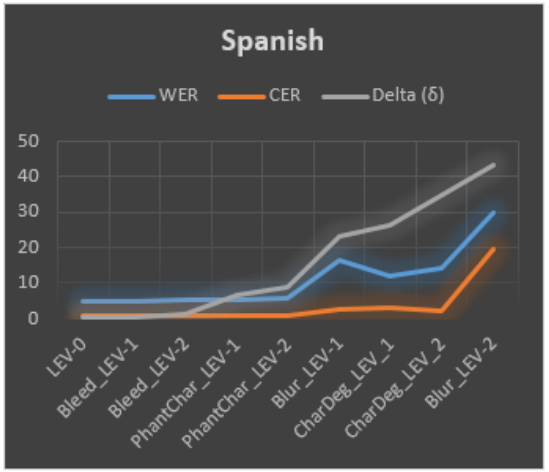
The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NER Evaluation (F1-score)

English	BLSTM-CRF	BLSTM-CNN	BLSTM-CRF-CNN	CoreNLP
Clean	90.17	90.77	**90.90**	85.10
LEV-0	86.77	86.93	87.45	79.61
Bleed_LEV-1	85.15	85.08	**86.11**	75.72
Bleed_LEV-2	84.63	**84.72**	83.96	75.27
Blur_LEV-1	**71.03**	70.99	**71.03**	63.39
Blur_LEV-2	59.77	58.98	**60.31**	49.15
DegChar_LEV-1	73.14	**74.22**	74.11	58.12
DegChar_LEV-2	**70.85**	69.43	68.77	55.06
PhantChar_LEV-1	85.59	85.67	**87.01**	74.21
PhantChar_LEV-2	84.58	85.03	**85.20**	73.66
LEV-MIX	**70.87**	70.11	70.82	63.35







The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Real-case data

Dataset

- OCRed NER corpus aligned with its GT provided by the National Library of Finland
- Language: Finnish
- Corpus: 450K tokens, 30K NEs
- Tagset: [PER, LOC]
- OCR error rates: CER = 7% ; WER = 17%
- Results:

		LOC	PER	TOT
clean	P	93.39%	87.43%	90.82%
	R	91.86%	84.68%	88.74%
	F1	92.62%	86.03%	89.77%
OCRed	P	89.68%	83.31%	86.97%
	R	91.06%	83.54%	87.83%
	F1	90.36%	83.42%	87.40%

Table 5. Results on the NLF corpus

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Conclusion and future work

Conclusion

- BLSTM models achieved satisfying results when OCR error rates are:
  - Less than ~15% at the character level
  - Less than ~30% at the word level
- Dataset of synthesized OCRed documents are made publicly available
- Results on synthetized documents are comparable to real-world documents
- Results provide guidance on the required OCR quality level for a targeted NER performance

Future work

- Deep analysis of OCR errors
- How NER approaches can overcome the OCR degradation and provide correct predictions?

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Thank you for your attention!

Ahmed Hamdi
ahmed.hamdi@univ-lr.fr