An Analysis of the Performance of Named Entity Recognition over OCRed Documents

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

firstname.lastname@univ-lr.fr

JCDL 2019

1. Context & Motivation

- In digital libraries, named entities are among the most relevant information to index documents and the main entry point to their retrieval.
- Most digitized documents are indexed through their OCRed version and OCR errors hinder their accessibility.
- Prior research shows that named entities are the first point of entry of DL users.

Goal: Quantitatively estimate the impact of OCR quality on NER performance.
Problem: No text with (original OCR) and (OCR groundtruth) and (annotated NER).
Methodology:

- Simulating OCR outputs from existing clean NER corpora with various levels and types of OCR noise.
- Testing state-of-the-art NER systems over clean and noisy data.
- Studying the correlation between NER results and OCR.

2. Overview on NER approaches

« *The* ***German*** *racing driver* ***Michael Schumacher*** *won with* ***Mercedes*** *the last race in* ***China****.* »

PER LOC ORG MSIC

Three main approaches:

- Rule-based: rules are based on dictionaries, triggers and linguistic descriptors
- Machine-learning: training, hand crafted features
- Deep-learning: jointly training and learning effective features

- → Rule-based systems are clearly disoriented to process noisy data: perpetual updates of rules for noisy NEs and costly manual efforts
- → Machine-learning systems can be updated and generalized.
- → Deep-learning systems outperform other machine-learning NER systems.

3. Dataset

- Starting from existing clean texts with NER GT (CoNLL-03 English corpus)
- Converting raw texts into images
- Synthetizing and injecting various types and levels of noise.
- Performing OCR over degraded images.

DocCreator to synthetize noise

- Character degradation: adding ink spots on characters
- Phantom character: adding phantom ink around characters.
- Bleeding effect: simulates the verso ink that seeps through the recto side
- Blurring effect

SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT .
Nadim Ladki
AL-AIN , United Arab Emirates 1996-12-06
Japan began the defence of their Asian Cup title with a lucky 2-1 win against Syria in a Group C championship match on Friday .
But China saw their luck desert them in the second match of the group , crashing to a surprise 2-0 defeat to newcomers Uzbekistan .

↓

SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT .
Nadim Ladki
AL-AIN , United Arab Emirates 1996-12-06
Japan began the defence of their Asian Cup title with a lucky 2-1 win against Syria in a Group C championship match on Friday .
But China saw their luck desert them in the second match of the group , crashing to a surprise 2-0 defeat to newcomers Uzbekistan .

↓

SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT .
Nadim Ladki
AL-AIN , United Arab Emirates 1996-12-06
Japan began the defence of their Asian Cup title with a lucky 2-1 win against Syria in a Group C championship match on Friday .
But China saw their luck desert them in the second match of the group , crashing to a surprise 2-0 defeat to newcomers Uzbekistan .

↓

SOCCER - JAPAN GE'F LUCKY WIN, CHINAtN SURPRISE DEFEAT .
Nadim Ladtgi'
AL—AIN, United ArabEmirates 1996-12-06,
.Japgfl beg'and'the defence of their Asian Cup title witha lucky 2-1 win against Syr'ia in a Grofip C championship. mat on ri ay. ,
_But China saw their luck desert them in the second match of the group , crasjnhg to a surprise 2-0 defeattanewcomers Uzbekistan . ' .

Character degradation

SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT .
Nadim Ladki
AL-AIN , United Arab Emirates 1996-12-06
Japan began the defence of their Asian Cup title with a lucky 2-1 win against Syria in a Group C championship match on Friday
But China saw their luck desert them in the second match of the group , crashing to a surprise 2-0 defeat to newcomers Uzbekistan

Phantom character

SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT .
Nadim Ladki
AL-AIN , United Arab Emirates 1996-12-06
Japan began the defence of their Asian Cup title with a lucky 2-1 win against Syria in a Group C championship match on Friday
But China saw their luck desert them in the second match of the group , crashing to a surprise 2-0 defeat to newcomers Uzbekistan .

Bleeding effect

SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT .
Nadim Ladki
AL-AIN , United Arab Emirates 1996-12-06
Japan began the defence of their Asian Cup title with a lucky 2-1 win against Syria in a Group C championship match on Friday
But China saw their luck desert them in the second match of the group , crashing to a surprise 2-0 defeat to newcomers Uzbekistan .

Blurring

SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT .
Nadim Ladki
AL-AIN , United Arab Emirates 1996-12-06
Japan began the defence of their Asian Cup title with a lucky 2-1 win against Syria in a Group C championship match on Friday
But China saw their luck desert them in the second match of the group , crashing to a surprise 2-0 defeat to newcomers Uzbekistan

RETAS to align OCR outputs and the GT

OCR : SOCCER - JAPAN GE'F@ LUCKY WI@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 wi@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

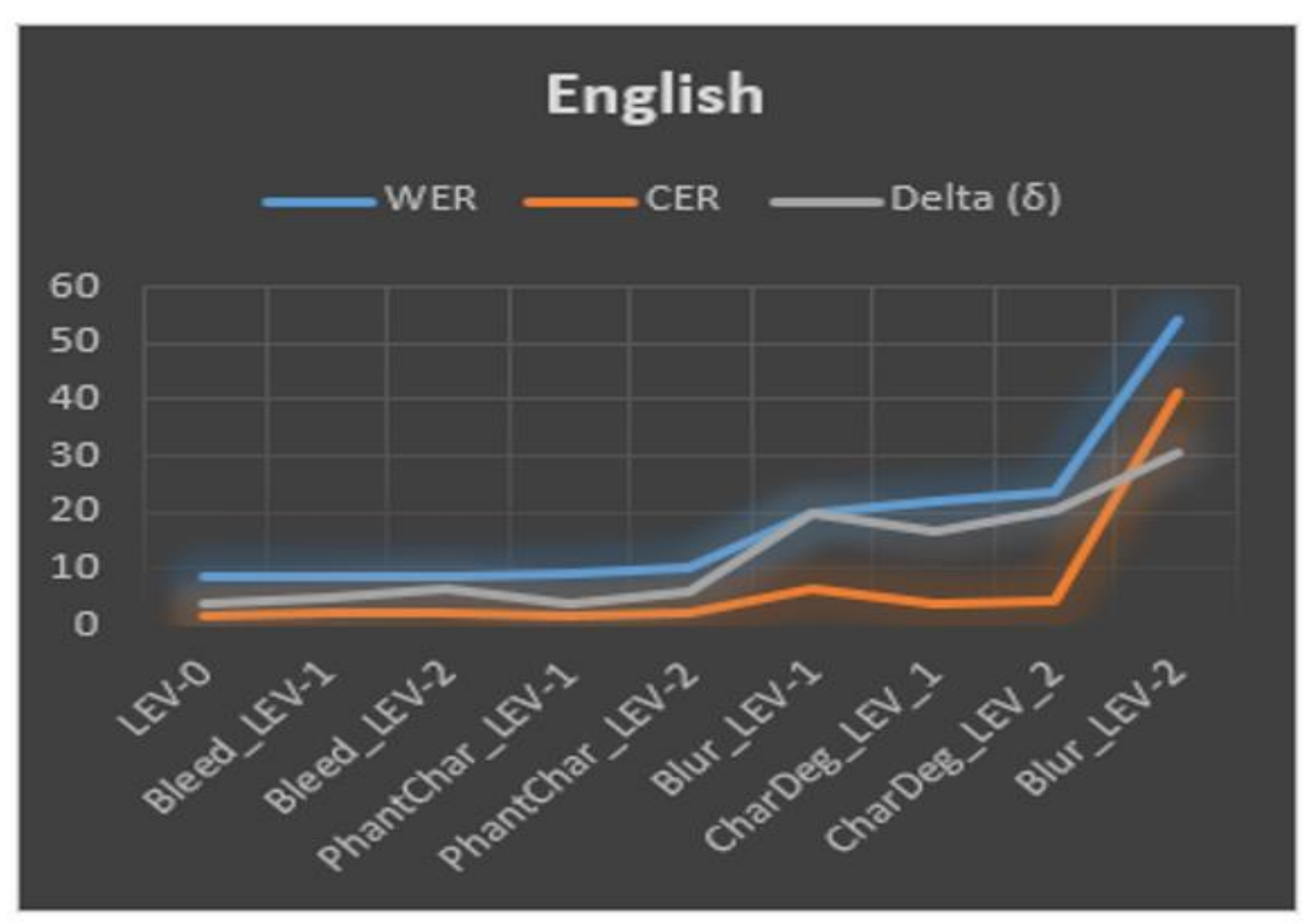
OCR : gainst Syr'ia in a Grofi@p C championship .mat@@ on @ri @ay@. , _ But China saw their luck desert the
GT : gainst Syr@ia in a Gro@up C championship @match on Fri@day . @@@@But China saw their luck desert the

4. Results & Evaluation

Neural-network based system: **LSTM-CRF**

GT:	Pierre	Van	Hoydonk	scored	for	Glasgow	Rangers
gold_annot:	PER	PER	PER	O	O	ORG	ORG
OCR lev-0:	Dierre	Van	Hoyoonk	scored	for	Glasgow	Rangers
pred_annot:	PER	PER	O	O	O	ORG	ORG
OCR blur:	Pierre	|n	Hoyoom	scores	on	Glasgow	Rimfen
pred_annot:	PER	O	O	O	O	LOC	O

	OCR			NER		
	CER	WER	ENER	Pre	Rec	F1-score
Clean	--	--	--	89.4	90.8	90.1
Lev-0	1.7	8.5	6.9	83.7	90.7	86.8
Bleed	1.8	8.6	7.1	84.0	84.1	84.1
Phantom-char	1.7	8.8	7.8	75.8	78.6	77.1
Blurring	6.3	20.0	21.5	66.9	69.5	68.8
Char-degradation	3.6	21.8	23.4	64,5	64.8	64.7



5. Conclusions & Future works

Conclusions

- Simulation of many OCR outputs similar to documents stored on digital libraries.
- Evaluation of the evolution of NER accuracy depending on the level of noise in the text.
- NER Accuracy drops from 90% to 60% with variable CER and WER from 1% to 7% and from 8% to 20% respectively.

Future works

- Adding weights to OCR outputs at the character level and the word level.
- NE-focused OCR post-correction => high impact on information access in digital libraries.

Acknowledgment

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant 770299 (NewsEye).