WHEN TO USE OCR POST-CORRECTION FOR NAMED ENTITY RECOGNITION?

ICADL 2020

30 November 2020

Vinh-Nam Huynh
Ahmed Hamdi
Antoine Doucet

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

Context

- Indexing documents stored in digital libraries requires OCR process.

- **Storage conditions, damaged documents** or **poor quality of printing materials** can lead to noisy texts strongly diverging from the original (i.e. the Ground Truth).

LA SITUATION POLITIQUE EN ALLEMAGNE
M. Marx va reconstituer
le cabinet Marx
Berlin, 28 mai. — Le président du Reich a chargé, ce matin, le chancelier démissionnaire, M. Marx, de la formation du cabinet. M. Marx a accepté cette mission. Devant le refus du leader du centre Stegerwald de former le nouveau cabinet et les réponses évasives du leader nationaliste Hergt, le président Ebert s'est décidé, comme on l-c prévoyait, à confier au chancelier démissionnaire le soin de reformer le cabinet.


LA SITUATION POLITIQUE EN ALLEMAGNE
M. Marx va reconstituer
le cabinet Marx
Berlin, 28 mai. — Le président du Reich a chargé, ce matin, le chancelier démissionnaire, M. Marx, de la formation du cabinet. M. Marx a accepté cette mission.
Devant le refus du leader du centre Stegerwald de former le nouveau cabinet et les réponses évasives du leader nationaliste Hergt, le président Ebert s'est décidé, comme on le prévoyait, à confier au chancelier démissionnaire le soin de reformer le cabinet.

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

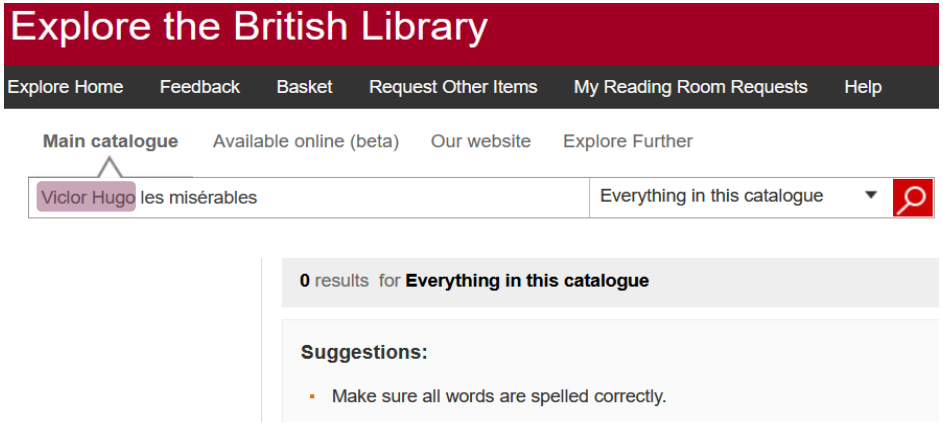
Named Entities and digital libraries

- NEs are the primary point of entry for users in a search system

➔ 80% of user queries on the Gallica digital library contain at least one named entity.

- **Named Entities (NEs)** are real-world objects that can be denoted with a proper name.

➔ It can have a physical existence or be abstract



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 important names and proper names in a given text and to categorise them into a set of predefined classes (person (PER). location (LOC). organisation (ORG). human product (PROD). etc.)

Paris Hilton stayed at the Hilton in Paris.

New York Times is based in New York.

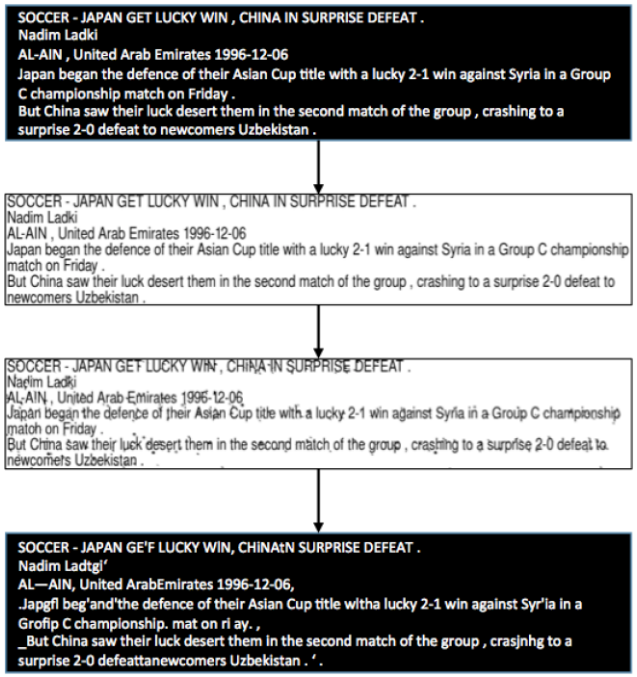
The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NER in noisy texts

Noisy data simulation

- No noisy NER corpora aligned with their clean versions
- Available clean NER data sets: CONLL-03
- Injection of OCR degradation
    - **Character degradation**
    - **Phantom degradation**
    - **Blurring**
    - **Bleeding effect**

https://zenodo.org/record/3877554#.XtmD_BY69uU



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 .

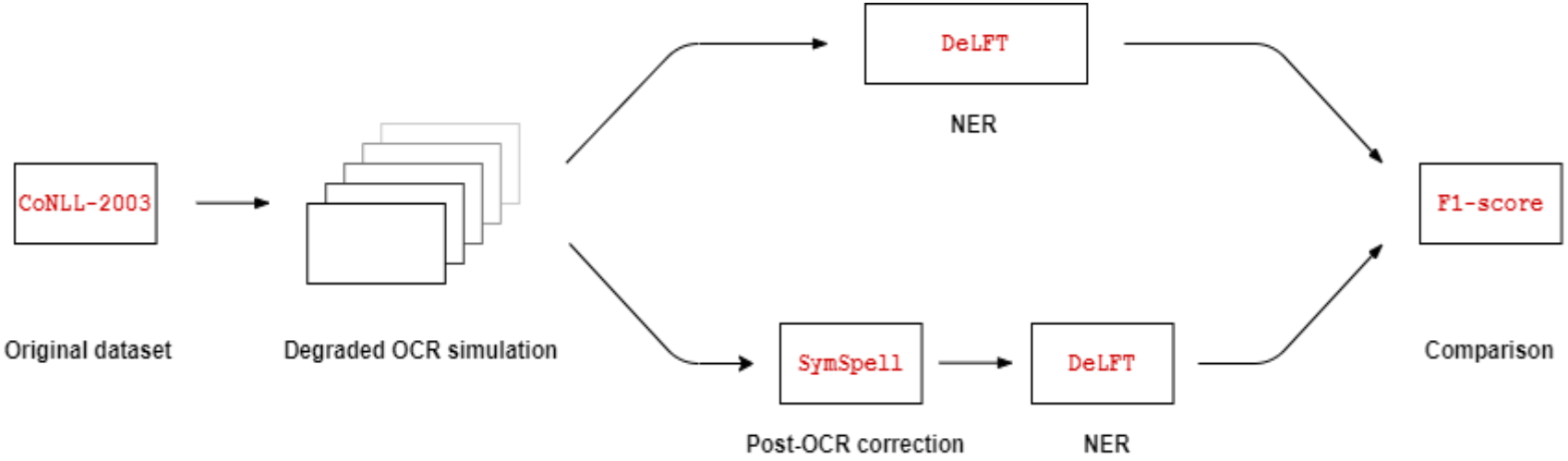
SOCCER - JAPAN GET LUCKY WIN , CHINA IN SURPRISE DEFEAT .
Nadim Ladki
AL-AIN , United Arab Emirates 1996-12-06
Japan began the defence of their Asian Cup title with a lucky 2-1 win against Syria in a Group C championship matoh on Friday .
But China saw their luck desert them in the second match of the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

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

When to use Post-OCR correction for NER?

Workflow



- NER systems: **BLSTM models** https://github.com/kermitt2/delft
- Post-OCR correction: **Symspell** https://github.com/mammothb/symspellpy

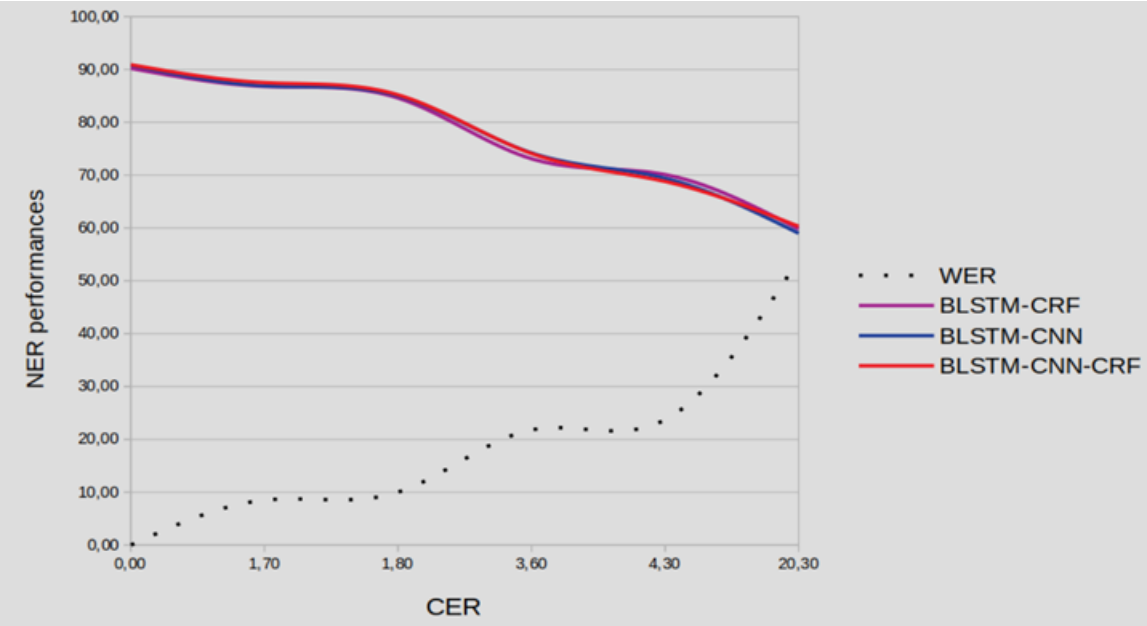
The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NER in noisy texts

Evaluation

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.

NER in noisy texts

Examples

- Noisy texts contain many out-of-vocabulary words.
  - Well recognised and classified contaminated NEs

    Clean: Mittermayer ➔ PER **Noisy**: Minermayer ➔ PER

  - Well recognised but bad classified contaminated NEs

    Clean: Charlton ➔ PER **Noisy**: Chalton ➔ ORG

  - Bad recognised and classified contaminated NEs

    Clean: Japan ➔ LOC **Noisy**: Japghl ➔ O

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

Post-OCR correction

SymSpell

- Publicly available on Github: https://github.com/wolfgarbe/SymSpell
- Three-steps processing:
  1. Error detection
  2. Candidate generation
  3. Filtering

OCR: Germany by the terms of the armistice is smpped

1- detection

2- candidate generation

stripped	0,92
skipped	0,05
snipped	0,01
slipped	0,01
.	
.	

3- filtering

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

Post-OCR correction

SymSpell

Performance:

- ~30,000 words / second (edit distance 2)
- ~50,000 words / second (edit distance 3)

Segmentation

```
- thequickbrownfoxjumpsoverthelazydog
+ the quick brown fox jumps over the lazy dog

- itwasabrightcolddayinaprilandtheclockswerestrikingthirteen
+ it was a bright cold day in april and the clocks were striking thirteen
```

Text correction

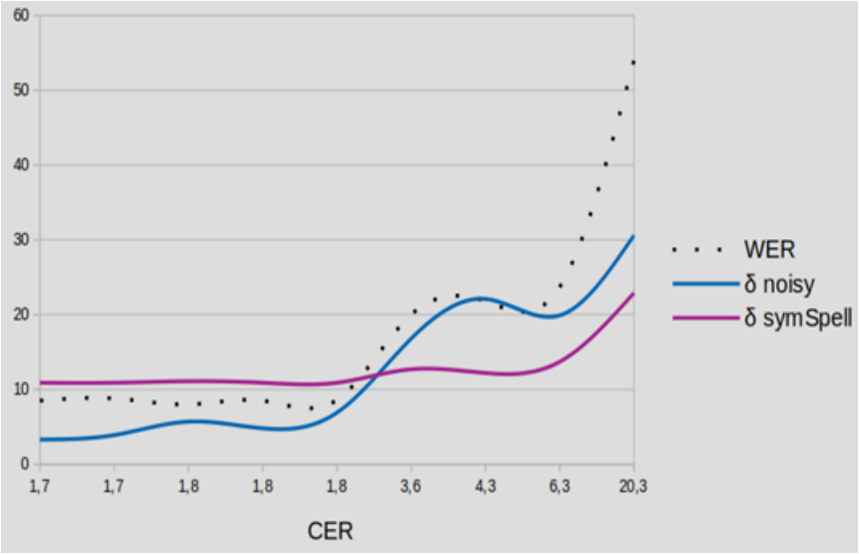
```
- in te dhird qarter oflast jear he hadlearned ofca sekretplan
+ in the third quarter of last year he had learned of a secret plan  (9 edits)

- the bigjest playrs in te strogsommer film slatew ith plety of funn
+ the biggest players in the strong summer film slate with plenty of fun  (9 edits)
```

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

Post-OCR correction for NER

Evaluation

OCR error rates		NER F1-score	
CER	WER	Original	SymSpell
CLEAN		90.4	--
1.7	8.5	**87.6**	80.0
1.7	8.8	**87.0**	80.0
1.8	8.0	**85.2**	79.8
1.8	8.5	**86.1**	80.0
1.8	8.6	**84.0**	80.0
3.6	20.0	74.1	**78.2**
4.3	21.8	68.8	**78.7**
6.3	23.7	71.0	**77.2**
20.3	54.0	60.3	**68.0**



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

Post-OCR correction for NER

Discussion

BEFORE

```
SOCCER - JAPAN GE'F LUCKY WlN, CHiNAtN SURPRISE DEFEAT .
Nadim Ladtgi
'AL—AIN, United ArabEmirates 1996-12-06, .
Japgfl beg'and'the defence of their Asian Cup title wltha lucky 2-1 win against Syr'ia in a Grofip C championship
```

AFTER

```
soccer - japan get lucky want china surprise defeat .
nadim ladki
al-ain, united arab emirates @ .
japan began @ defence of their asian cup title with lucky 21 win against syria in a group c championship
```

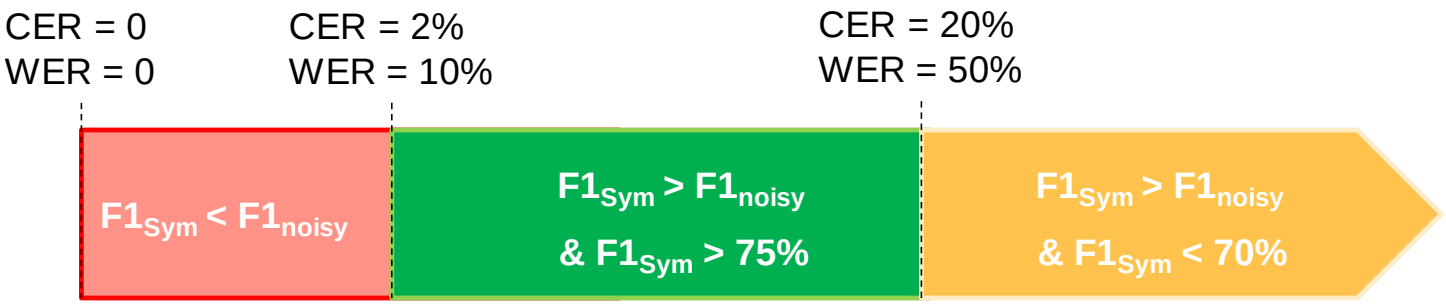
Pros: when WER < **25**%, NER F1-score is boosted up to **77** %

Cons: when WER < **10**%, post-OCR may degrade NER F1-score

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

Conclusions

- The SymSpell algorithm consistently increases NER results over noisy texts when the CER and the WER respectively exceed 2% and 10%.



- For future works:
  1. deep analysis on the impact of OCR quality on NER
  2. what about other post-OCR correction techniques?

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

Questions?

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