Distant Supervision and Noisy Label Learning for Low Resource Named Entity Recognition: A Study on Hausa and Yorùbá

David I. Adelani*, Michael A. Hedderich*, Dawei Zhu*,

Esther van den Berg and Dietrich Klakow

PML4DC @ ICLR 2020

SIC Saarland Informatics Campus

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Named Entity Recognition (NER)

- Core NLP task
- Recognizing entities like Person, Location, Organization or Date
- Classification task

On	O
the	O
4th	B-DATE
of	I-DATE
February	I-DATE
,	O
Global	B-ORG
Voices	I-ORG
visited	O
Fernando	B-PER
Gomes	I-PER

Named Entity Recognition (NER)

- Core NLP task
- Recognizing entities like Person, Location, Organization or Date
- Classification task

- High resource settings ≈ 90 F1 score
- Many African languages lower

On	O
the	O
4th	B-DATE
of	I-DATE
February	I-DATE
,	O
Global	B-ORG
Voices	I-ORG
visited	O
Fernando	B-PER
Gomes	I-PER

Named Entity Recognition (NER)

- Core NLP task
- Recognizing entities like Person, Location, Organization or Date
- Classification task

- High resource settings ≈ 90 F1 score
- Many African languages lower

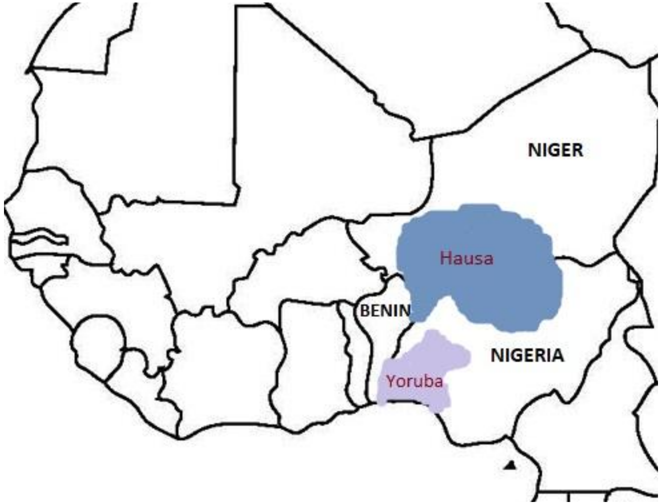
- Low-resource settings
  - Pretrained word embeddings
  - Distant supervision
  - Label-noise handling

On	O
the	O
4th	B-DATE
of	I-DATE
February	I-DATE
,	O
Global	B-ORG
Voices	I-ORG
visited	O
Fernando	B-PER
Gomes	I-PER

Two African Languages

- Hausa & Yorùbá
  - Second and third most spoken *indigenous language*
  - Over 40 million and 35 million native speakers, resp.

- Hausa NER data
  - LORELEI language pack [Strassel & Tracey, LREC 2016]
  - Used in collaboration with CMU
- Yorùbá NER data
  - Global Voices news articles [Alabi et al., LREC 2020]



Locations of native speakers

Pretrained Word Embeddings & Models

- Vector representation for words
- Pretrained on unlabeled text

- FastText [Bojanowski et al., arxiv 2016] + Bi-LSTM
  - Word representation based on subwords
  - Different Yorùbá sources (incl. JW300, News, Wikipedia, Twitter) [Alabi et al., LREC 2020]
  - 1.5M parameters

- BERT [Devlin et al., NAACL 2019] + CRF
  - Contextual word embeddings
  - Multilingual model on Wikipedia
  - 110M parameters

Training Data through Distant Supervision

Clean, expensive, manually-annotated text

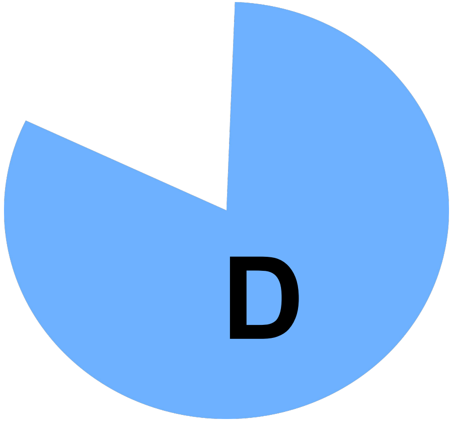
Unlabeled text

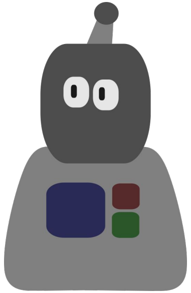
Training Data through Distant Supervision

Clean, expensive,
manually-annotated text

Unlabeled text

+ automatic annotation
(quick + cheap)



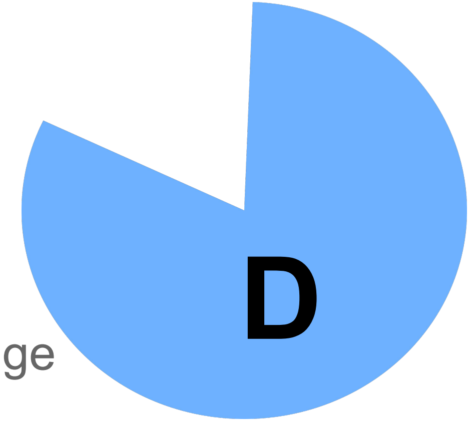


Training Data through Distant Supervision

Clean, expensive,
manually-annotated text

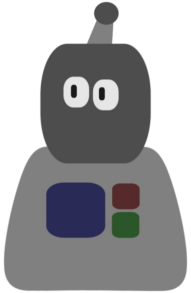
Unlabeled text

+ automatic annotation (quick + cheap)



Leverage

- context
- expert insights
- external knowledge and resources
- self-training



Distant Supervision

Rules

- Native speaker (domain expert)
- Date detection using keywords like "ọjọ́" (day) "oṣù" (month)

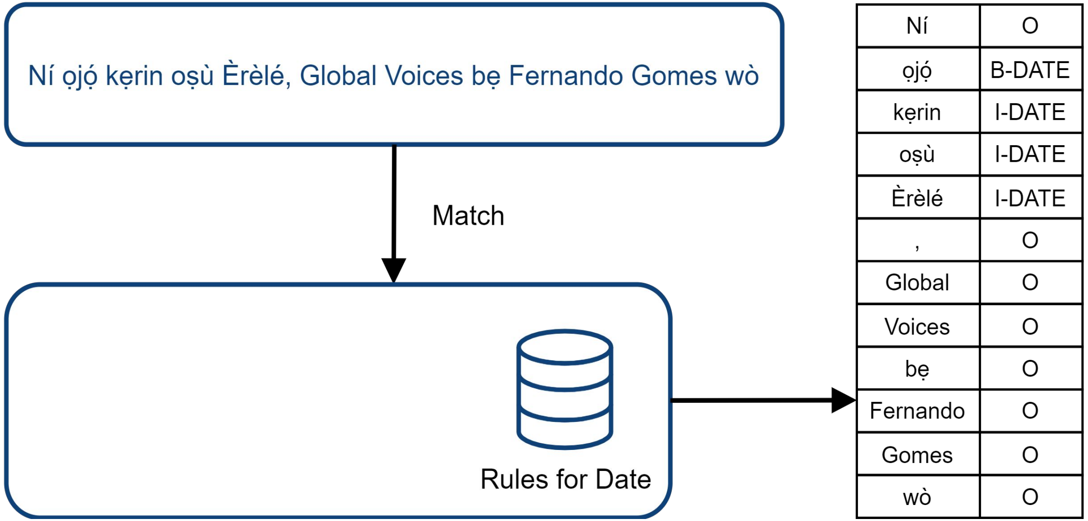
Ní ọjọ́ kẹrin oṣù Èrèlé, Global Voices bẹ Fernando Gomes wò

Ní	O
ọjọ́	O
kẹrin	O
oṣù	O
Èrèlé	O
,	O
Global	O
Voices	O
bẹ	O
Fernando	O
Gomes	O
wò	O

Distant Supervision

Rules

- Native speaker (domain expert)
- Date detection using keywords like "ọjọ́" (day) "oṣù" (month)



Ní ọjọ́ kẹrin oṣù Èrèlé, Global Voices bẹ Fernando Gomes wò

Ní	O
ọjọ́	B-DATE
kẹrin	I-DATE
oṣù	I-DATE
Èrèlé	I-DATE
,	O
Global	O
Voices	O
bẹ	O
Fernando	O
Gomes	O
wò	O

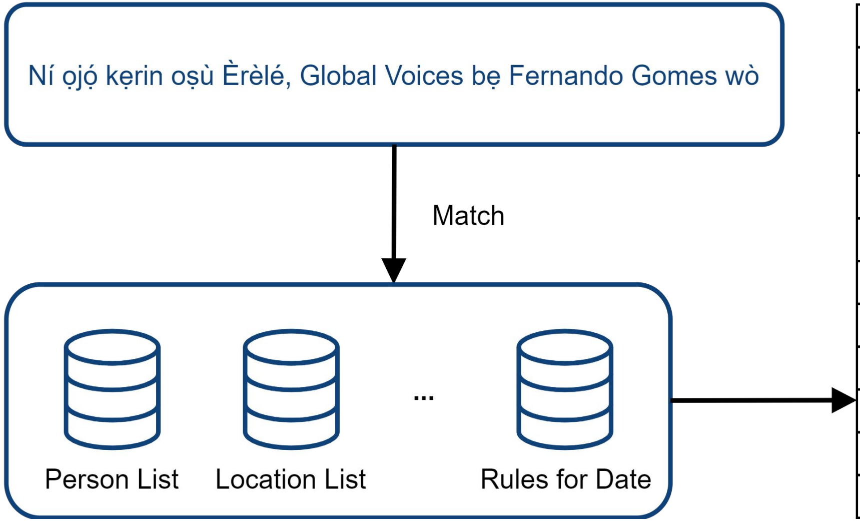
Distant Supervision

Rules

- Native speaker (domain expert)
- Date detection using keywords like "ọjọ́" (day) "oṣù" (month)

Entity lists

- From sources like gazetteers, dictionaries, phone books, and Wikipedia



Ní ọjọ́ kẹrin oṣù Èrèlé, Global Voices bẹ Fernando Gomes wò

Ní	O
ọjọ́	B-DATE
kẹrin	I-DATE
oṣù	I-DATE
Èrèlé	I-DATE
,	O
Global	O
Voices	O
bẹ	O
Fernando	B-PER
Gomes	I-PER
wò	O

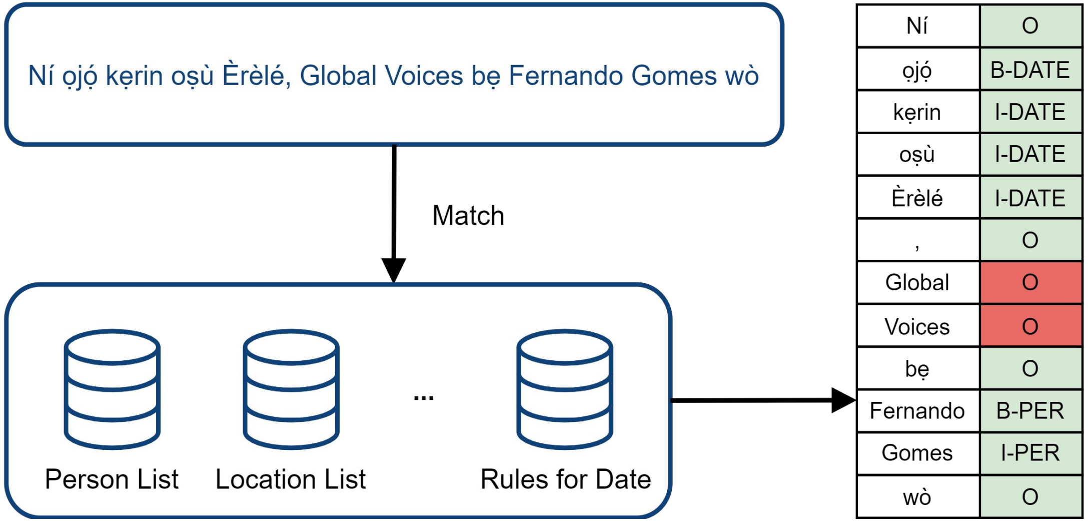
Distant Supervision

Rules

- Native speaker (domain expert)
- Date detection using keywords like "ọjọ́" (day) "oṣù" (month)

Entity lists

- From sources like gazetteers, dictionaries, phone books, and Wikipedia



Ní ọjọ́ kẹrin oṣù Èrèlé, Global Voices bẹ Fernando Gomes wò

Ní	O
ọjọ́	B-DATE
kẹrin	I-DATE
oṣù	I-DATE
Èrèlé	I-DATE
,	O
Global	O
Voices	O
bẹ	O
Fernando	B-PER
Gomes	I-PER
wò	O

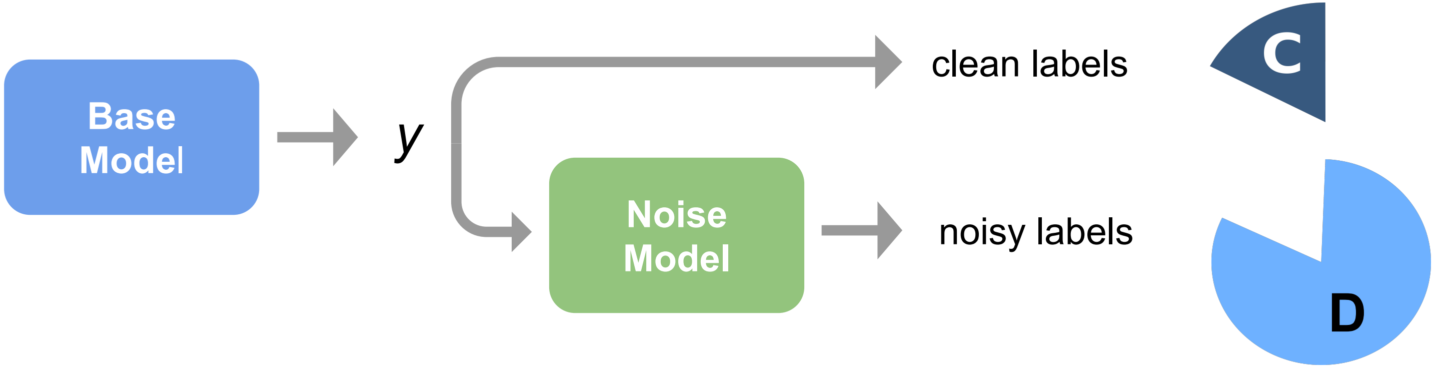
Label-noise handling

- Distant supervision usually more errors ⟶ noisy labels
- Can deteriorate performance
- Explicit noise handling
  - Noise modeling
  - Label cleaning

Named Entity class	F1-score
Overall	41
PER	22
LOC	62
ORG	22
DATE	48

Quality of Distant Supervision

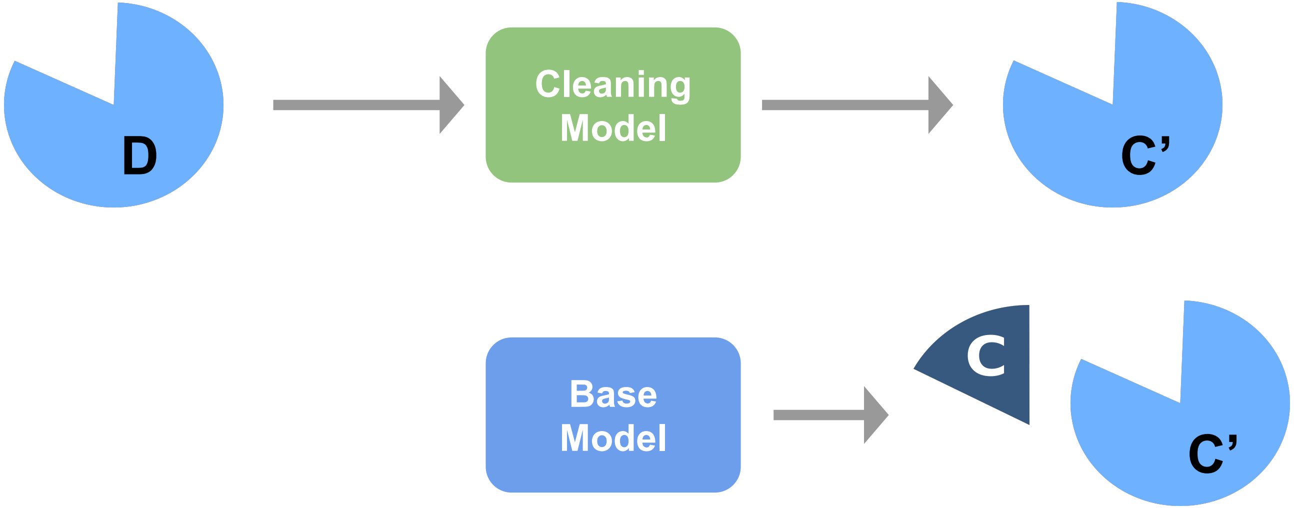
Noise Modeling



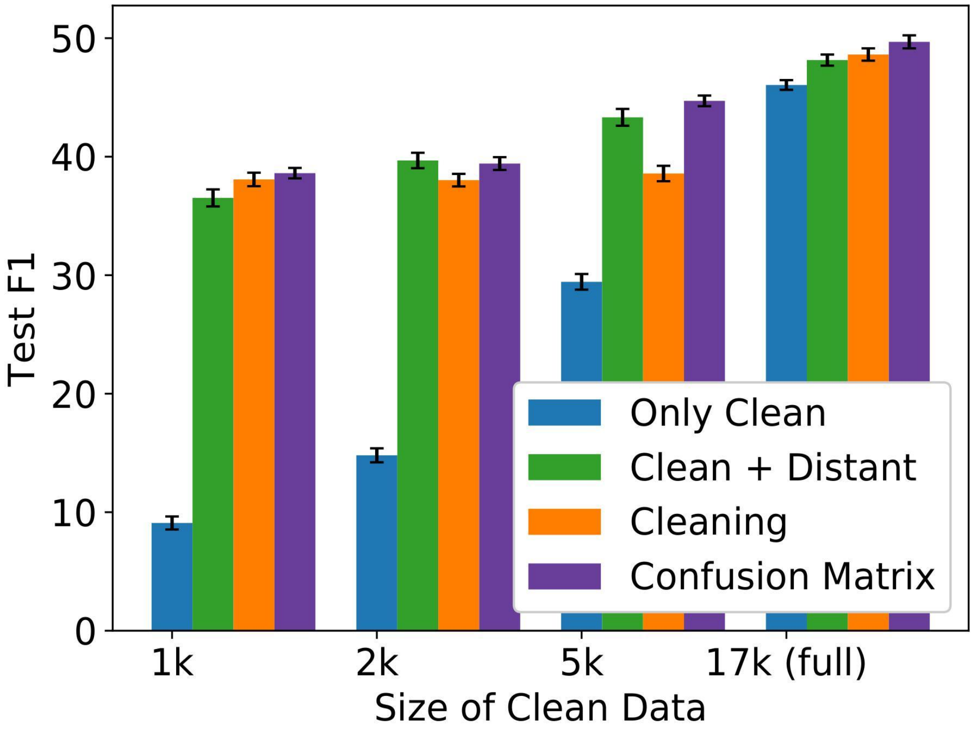
- Noise Channel [Bekker & Goldberger, ICASSP 2016]
  - EM algorithm
- Confusion Matrix [Hedderich & Klakow, DeepLo 2018]
  - Pairs of clean and noisy labels

Noise Cleaning

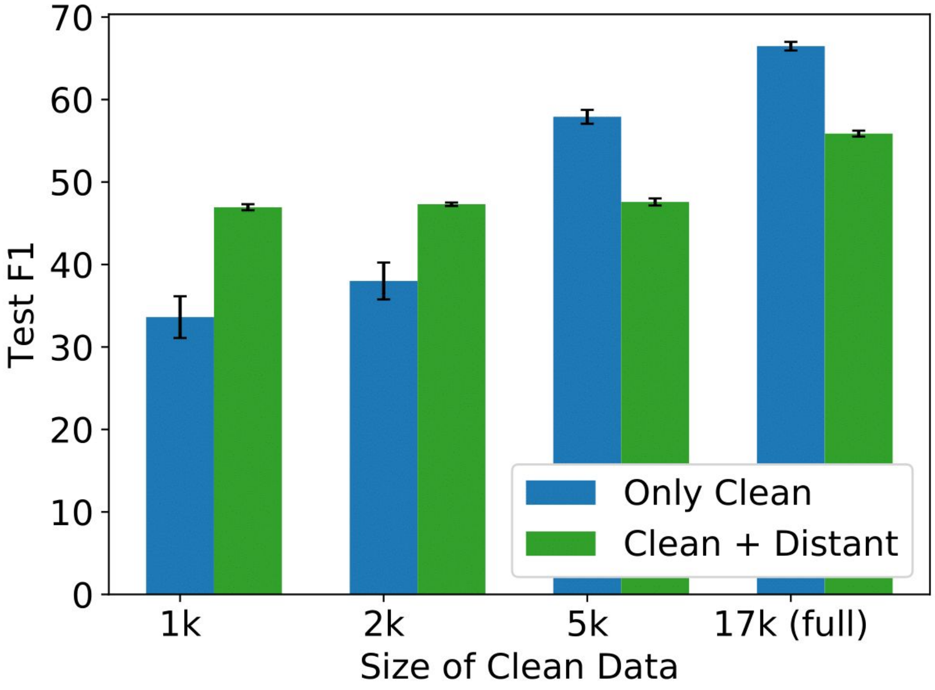
[Veit et al., CVPR 2017]



Results on Yorùbá: FastText + Bi-LSTM



Results on Yorùbá: BERT + CRF



Summary

- NER in Low-Resource Settings
- Pretrained word embeddings
  - Trade-off model size and performance
- Distant supervision
  - Can boost performance through automatically obtained labels
- Label-noise handling
  - Reduce negative effects of noise in distant supervision

More details + experiments in the paper:
https://arxiv.org/abs/2003.08370

{didelani,mhedderich,dzhu}@lsv.uni-saarland.de