Post-OCR Error Detection and Correction

Through the sociopolitical lens of Sivaji: The Boss (2007)[1]

Presented by Surya

Collaborators: Sharmila, Sumit, Balaji, Aparna

[1] *Sivaji: The Boss*. Directed by Shankar, performances by Rajinikanth and Suman, AVM Productions, 2007

OCR Technology

- Acquire vast information in documents – digitize, search, retrieve, summarize
- Fast, cheap, and secure compared to human annotators

- Lots of errors in OCRed text leading to poor downstream task performance
- NER, Coreference Resolution, POS Tagging – all plagued by low accuracies

Why do OCR Errors occur?

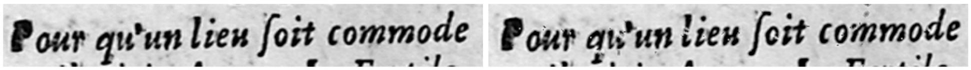
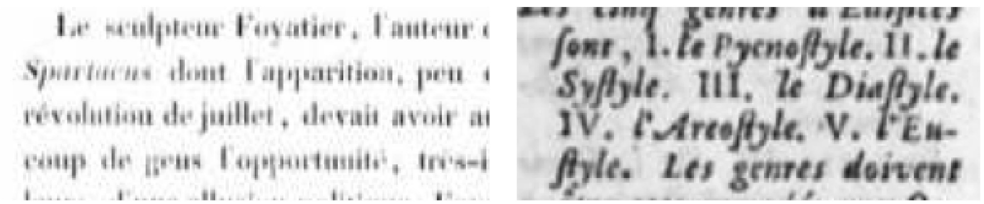
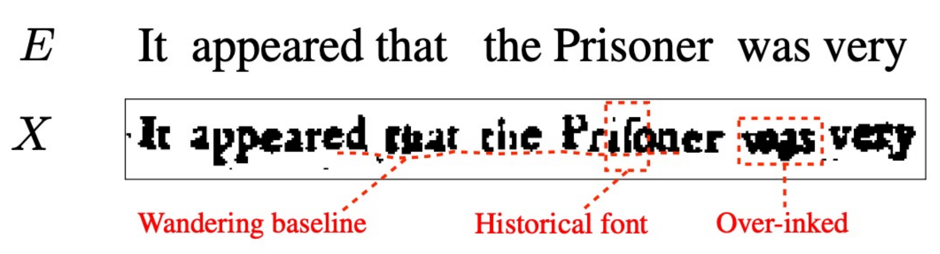


Figure 3. Ink degradation on an old document. (**Left**) original image. (**Right**) degraded image.

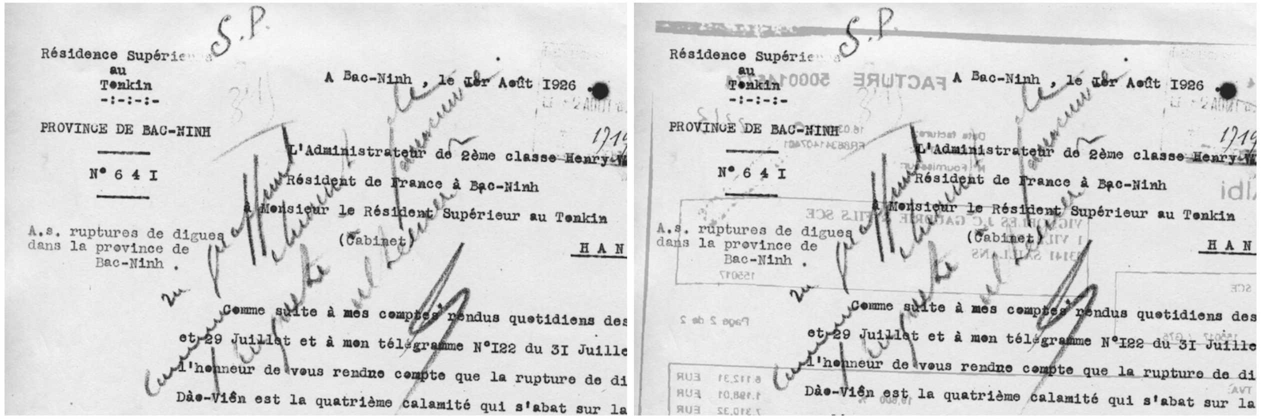
Ink Degradation: Small ink spots on characters;
Due to old docs or poor scanning



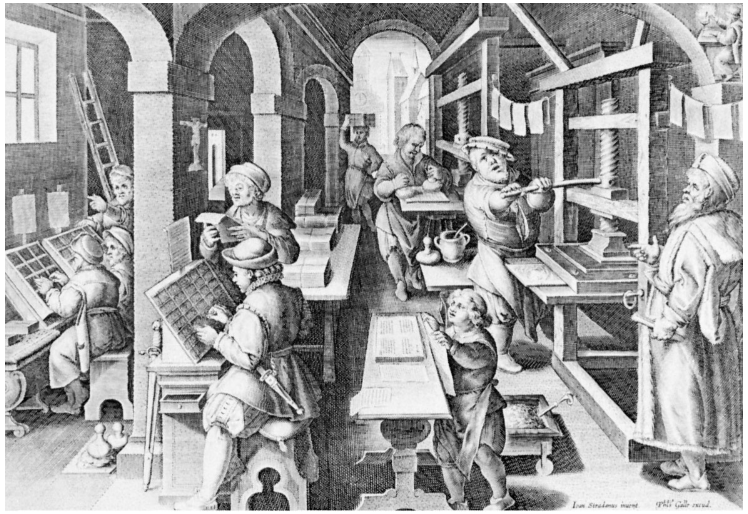
Blurring



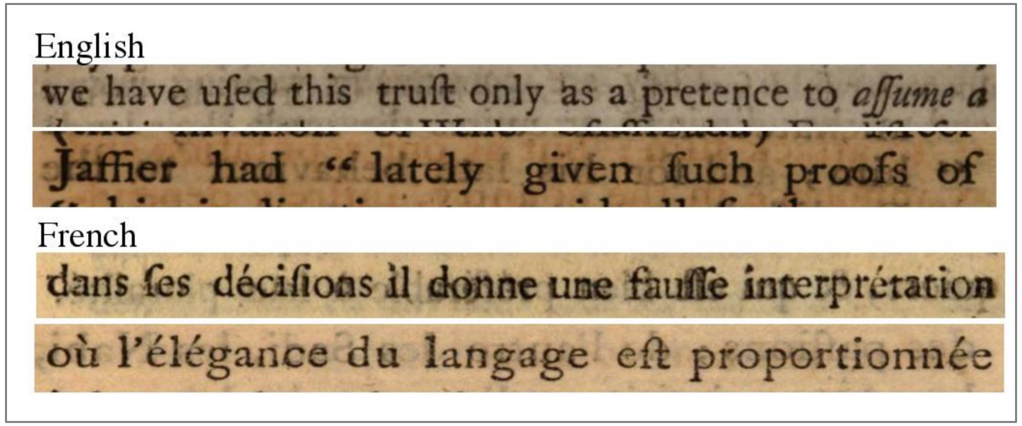
Historical Fonts; Ink Spots, Bad wooden printing machines



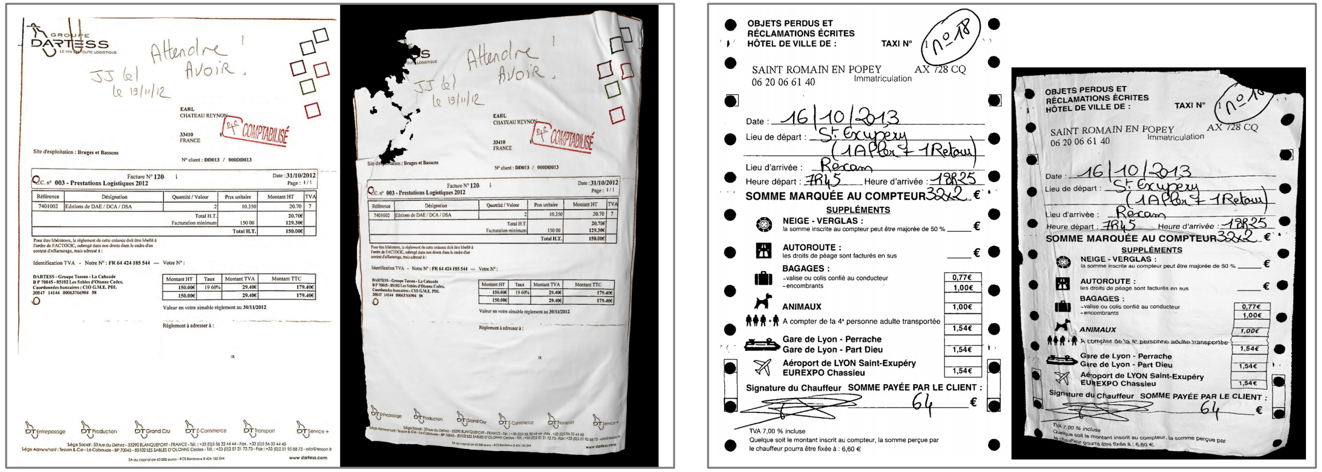
Bleedthrough Effect: you can see text from previous pages
Handwriting on docs



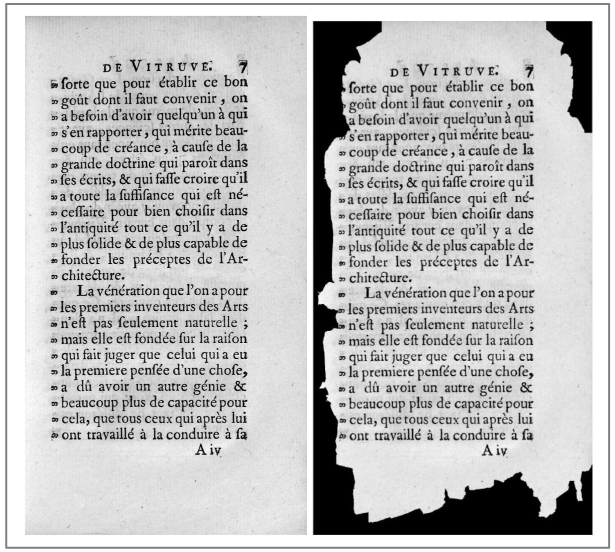
Why do OCR Errors occur?



Non-English languages

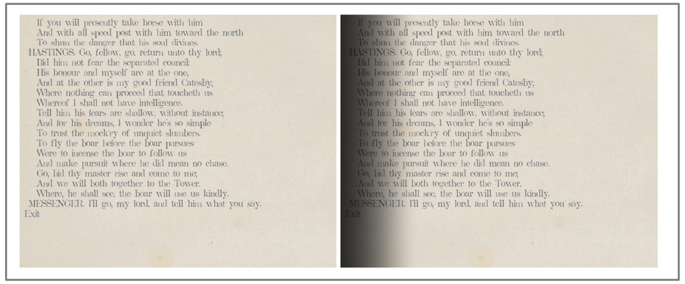


3D Deformations due to Poor scanning



Torn / Burned Pages

Images clicked at bad viewpoint



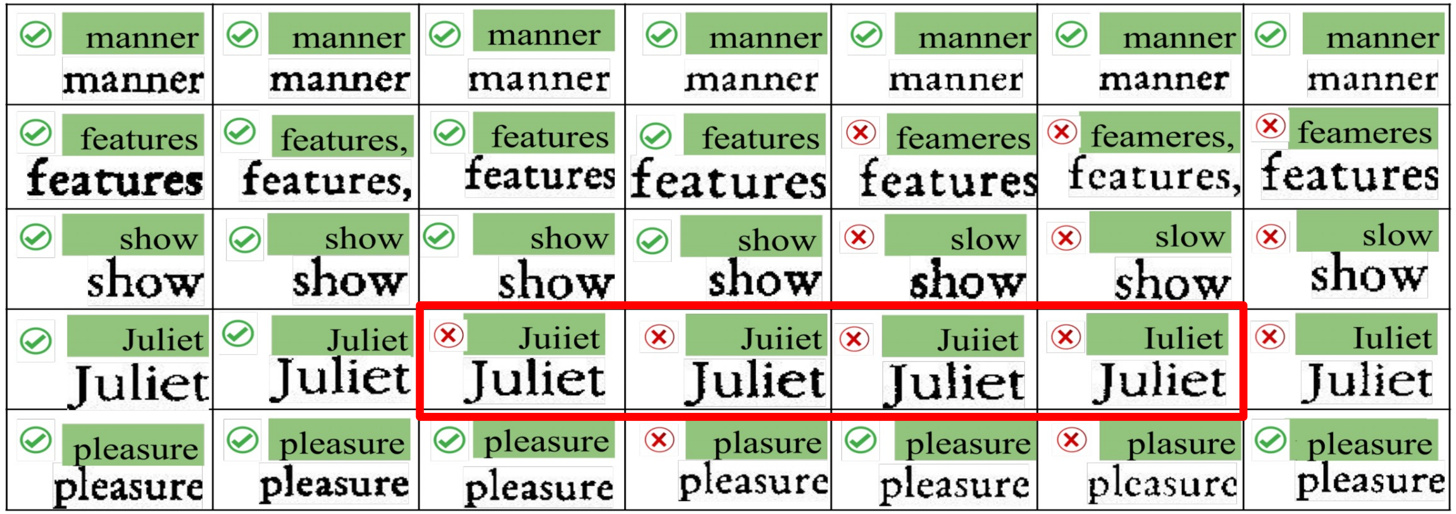
Poor Illumination

Impact of OCR Errors

As CER increases to **6%**, NER F1 Score drops by **25 points**

	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
PhantChar	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
CharDeg	3.6	21.8	23.4	64.5	64.8	64.7

Table 1: NER performance over noisy data, for undegraded OCR (LEV-0), bleed-through (Bleed), phantom degradation (PhantChar), Blurring effect and character degradation (CharDeg)



- Heavily dependent on vision / perceptual data
- Do not take semantics / words into account
- However, the errors are somewhat repetitive
  - "li" --> "ii" ; "J" --> "I"

=> There seems to be some structure

- Improving OCR itself is one way to tackle
  - But not all have **access** to AWS, GCP, Azure
  - Or the **original images**
- **Easy to iterate** on a post-processing model that can be finetuned for your datasets

SUPER
BERTji
STAR

```
from transformers import BertModel
```

BERT Arrives!

- Visual Euphemism of large autoregressive LMs such as BERT, RoBERTa, GPT, etc.

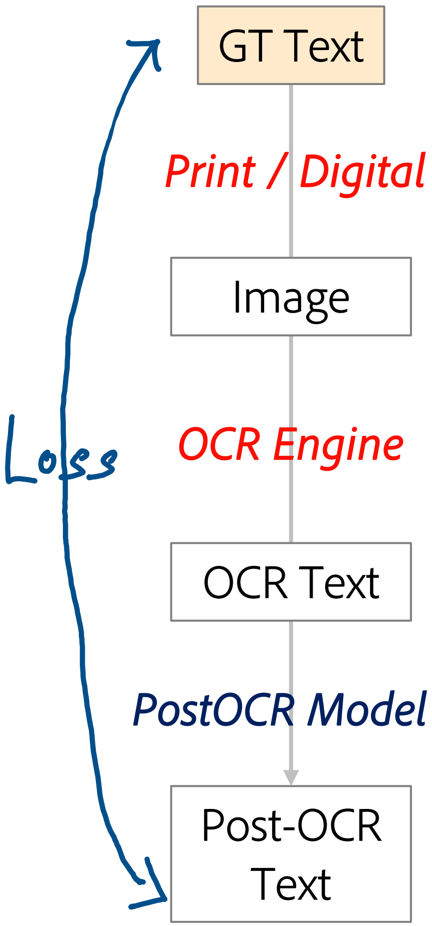
$$p(x_{1:T}) = \prod_{t=1}^{T} p(x_t | x_{<t})$$

- Great performance on many NLP benchmarks such as GLUE, SuperGLUE

- BERTology and other interpretability papers have found good meaning representation in the vector space

- Can impart context + word-meaning to OCRed text; potentially helping in correction

Problem Formulation



- Easily available data = OCR Text (+ related images)
- Tough to obtain = Clean Text

- Tough to simulate errors – will discuss error types later

Given a sequence of *n* OCR tokens **S**, the objective is to find true word sequence **W** that is printed in the original text.

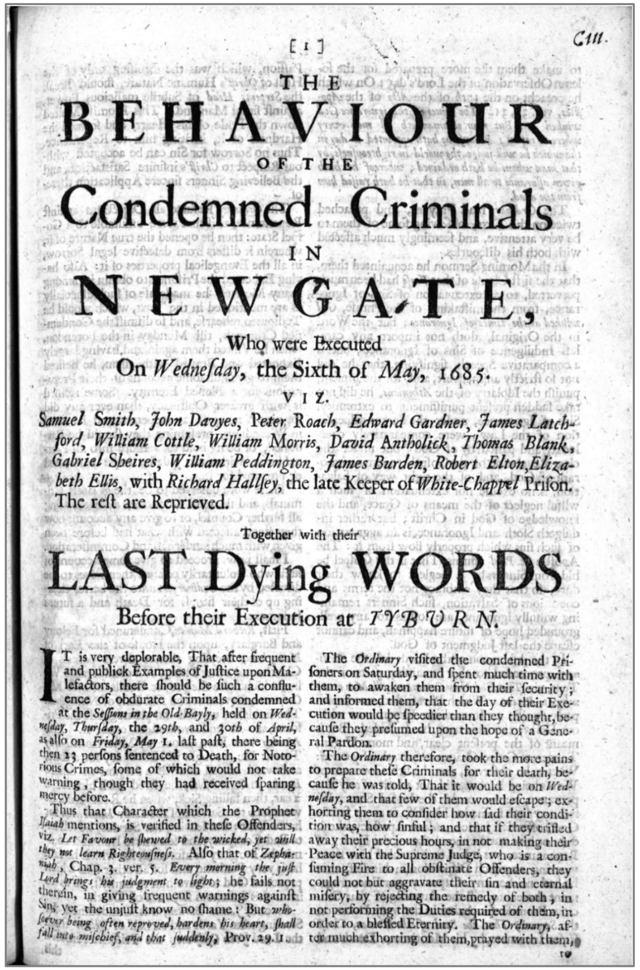
$$S = [s_1 \ s_2 \ \ldots \ s_n]$$

$$W = [w_1 \ w_2 \ \ldots \ w_m]$$

$$\hat{W} = \arg\max_{W} P(S|W)\,P(W)$$

$$Loss = F(W, \hat{W})$$

Tough to Align


[1]

THE
BEHAVIOUR
OF THE
Condemned Criminals
IN
NEWGATE,
Who were Executed
On Wednesday, the Sixth of May, 1685.
VIZ.
Samuel Smith, John Davyes, Peter Roach, Edward Gardner, James Latchford, William Cottle, William Morris, David Antholick, Thomas Blank, Gabriel Sheires, William Peddington, James Burden, Robert Elton, Elizabeth Ellis, with Richard Hallsey, the late Keeper of White-Chappel Prison.
The rest are Reprieved.
Together with their
LAST Dying WORDS
Before their Execution at TYBURN.

OCR Text

```
E i
CIIL.
T H E
B
O
F T H E
Condemned
Criminals
I N
NE
W
G
Who were Executed
On Wednefday, the Sixth of May, I 685.
V I Z.
Samuel Smith, Fobn Dauyes, Peter Roach, Edward Gardner, Fames Lat
Gabriel ford, Wiliam Sbeires, Cottle, Wiliam William Peddington,
beth Elis, with Richard Hallfey, the late Keeper of White-Chappel
The ret are Reprieved.
Together with their
LAST Dying WORD
Before their Execution at TYB 7 R N.
yllothw
T is very deplorable, That after frequent
The Ordinary vifited the condemned Pei-
and publick Examples of Jufice upon Ma-
foners on Saturday, and fpent much time with
lefaators, there should be fuch a conflu+
them, to awaken them from their fecurity.;
ence of obdurate Criminals condemned
and informed them, that the day of their Exe+
at the Selions in the Old:Bayly, held on Wed.
cution would be fpeedier than they thought, be-
nefday, Thurfday, the 29th, and 3oth of April,
caufe they prefumed upon the hope of a Gene-
alfo on Friday, May I. laft paft, there being
ral Pardon.
then 23 perfons fentenced to Death, for Noto
The Ordinary therefore, took the mote pains
rious Crimes, fome of which would not take
to prepare thefe Criminals for their death; be-
warning 5 though they had received fparing
caufe he was told, hat it would be on Wed-
mercy before.
nefday, and that few of them would efcapel; EXH
Thus that Character which the Prophet
horting them to confider how fad their condi-
Ifaiab mentions, is verified in thefe Offenders,
```

GT Text

```
THE BEHAVIOUR OF THE Condemned Criminals IN  NEWGATE ,
 Who were Executed On Wednesday, theSixth May 1685.
 VIZ. Samuel Smith, John Davyes, Peter Roach, Edward Gard
Richard Hallsey, the late Keeper of White-Chappel Prison.
 Together with their LAST Dying WORDS Before their Execut
 IT is very deplorable, That after frequent and publick E
April, as also on Friday May 1. last past, there being th
 Thus that Character which the Prophet Isaiah mentions, i
just Lord brings his judgment to light; he fails not ther
suddenly, Prov. 29. I.
 The Ordinary visited the condemned Prisoners on Saturday
prefumed upon the hope of a General Pardon.
 The Ordinary therefore, took the more pains to prepare t
was, how sinful; and that if they trifled away their prec
by rejecting the remedy of both, in not performing the Du
```

Old Bailey Dataset with GT Text and OCR Text (+ related images) but no alignment

Types of Errors

```
({None: 367119,
  'Misrecognition': 2514,
  'ExtraContent': 9002,
  'ContentLoss': 2615,
  'UnderScore': 654,
  'RemovedSpacing': 1008,
  'Punctuation': 738,
  'Hyphenation': 27,
  'CapsError': 37,
  'Shapes': 529,
  'ExtraSpacing': 2},
```

Non word error – easy
“ant” -> “amt”

Real word error – tough; requires context
“ant” -> “aunt”

- 1k documents ; 10k sentences
- All errors are word level

- Misrecognition (“main” -> “rnain”)
- Extra Content (“he re” -> “here”)
- Content Loss (“”)
- Hyphenation (URLs, linebreaks)

- Very hard to simulate from clean text data
  - OCR Engine dependent artifacts
  - Dataset/page dependent artifacts
  - They aren’t just any ED-1 errors
- “scho ol” -> “school”; “sch ool” -> “school”
  - Can’t create a relation between length of word and probability of OCR error

- Computing such stats requires 100% alignment
- A better way to simulate is to generate noisy images and **pass through the OCR engine**

Previous Attempts

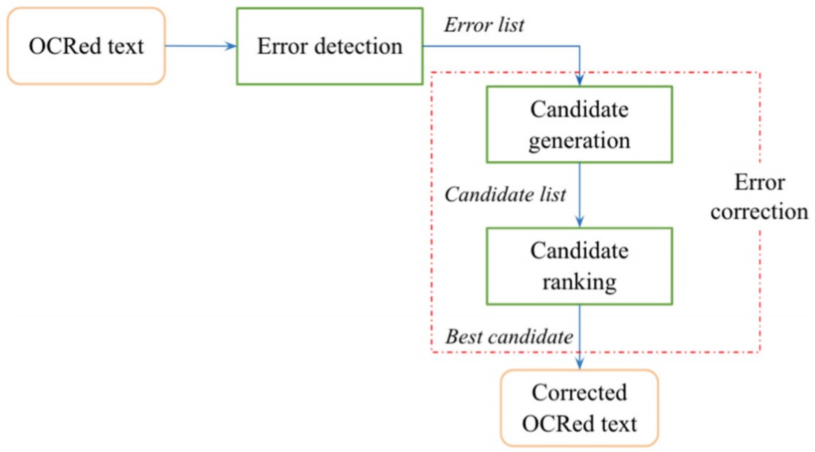


Fig. 1. Post-OCR processing pipeline.

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X \ Y (correct)	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

Whitespace Error Correction technique

- **Candidate Generation**: Consider all possible splits
- **Candidate Ranking**: Find the split that suits the scenario well
- Exponential complexity; although simplified with some assumptions

Token Dictionary Similarity

- For each token, replace with the closet high frequency token in dictionary that suits the context well
- Doesn't work for real-word errors

Spellcheckers

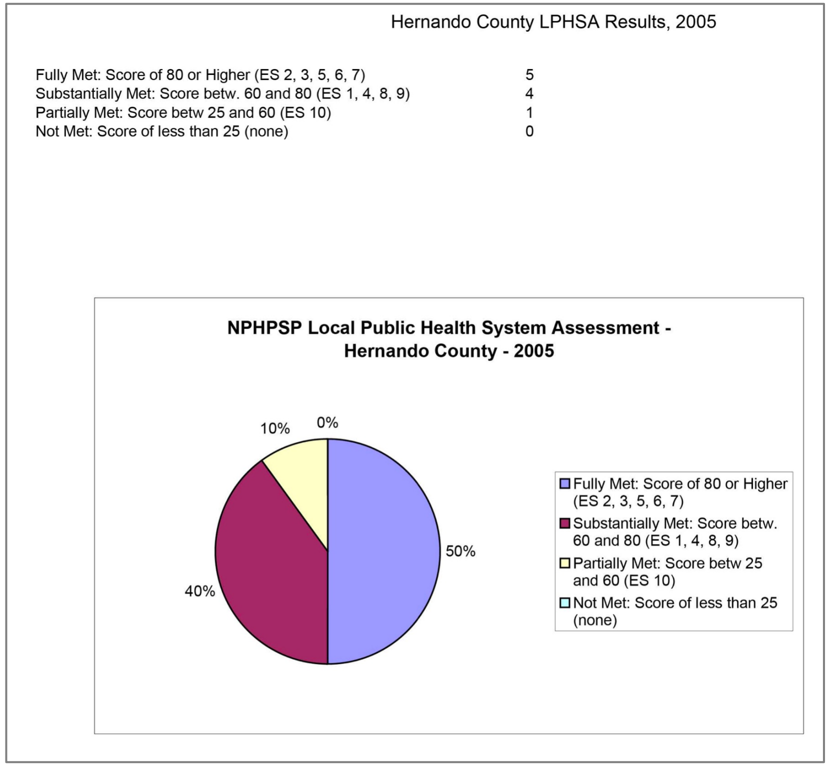
- Character Confusion Matrix
- Neural methods (NeuSpell)
- Errors here are not necessarily spelling mistakes; different structure

Tough PDFs


```
004340.pdf
V #1es of Insurance#Related Fraud # @@@@@@@@@ @@@@@ @@@@@@ @ @@@@@@@@@@@@ @@@@@@@ @@ @@@@@@@ @@@@@@@@ L 1uisia no # o u #
Theft for salvage # Cleaning and resale elsewhere # Fraud # Multiple claims for preexisting damage # Claims for damage not
caused by disaster # Phony #forged receipts for personal property loss# hotel stays # Phony insurance aqjuster#direct bill
ing to victims for poor or incomplete repair work
416
-------------
Varieties of Insurance# Related Fraud # Louisiana State Police # Unauthorized removal of flooded vehicles # Theft for @ @
@ salvage # Cleaning and resale elsewhere # Fraud # Multiple claims for preexisting damage # Claims for damage not caused
by disaster # Phony# forged receipts for personal property loss# hotel stays # Phony insurance adjuster# direct billing to
victims for poor or incomplete repair work
409
```

Some pages contain footnotes, sidenotes, etc. and alignment becomes very tough without GT guidance

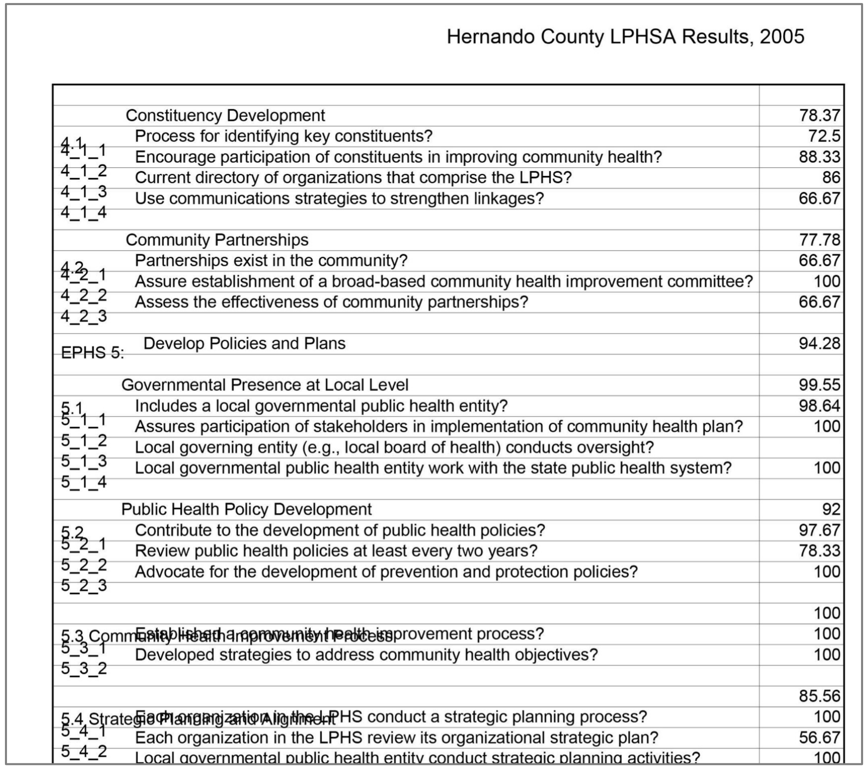
Tougher PDFs



Hernando County LPHSA Results, 2005

Fully Met: Score of 80 or Higher (ES 2, 3, 5, 6, 7)	5
Substantially Met: Score betw. 60 and 80 (ES 1, 4, 8, 9)	4
Partially Met: Score betw 25 and 60 (ES 10)	1
Not Met: Score of less than 25 (none)	0

PDFs with Figures – text within figures



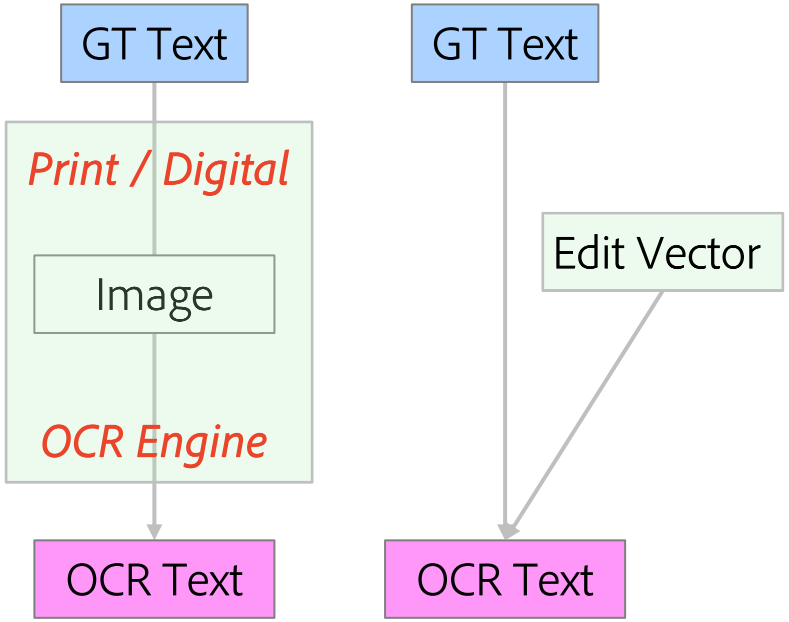
Hernando County LPHSA Results, 2005

	Constituency Development	78.37
4.1 4_1_1	Process for identifying key constituents?	72.5
4_1_2	Encourage participation of constituents in improving community health?	88.33
4_1_3	Current directory of organizations that comprise the LPHS?	86
4_1_4	Use communications strategies to strengthen linkages?	66.67
	Community Partnerships	77.78
4.2 4_2_1	Partnerships exist in the community?	66.67
4_2_2	Assure establishment of a broad-based community health improvement committee?	100
4_2_3	Assess the effectiveness of community partnerships?	66.67
EPHS 5:	Develop Policies and Plans	94.28
	Governmental Presence at Local Level	99.55
5.1 5_1_1	Includes a local governmental public health entity?	98.64
5_1_2	Assures participation of stakeholders in implementation of community health plan?	100
5_1_3	Local governing entity (e.g., local board of health) conducts oversight?	
5_1_4	Local governmental public health entity work with the state public health system?	100
	Public Health Policy Development	92
5.2 5_2_1	Contribute to the development of public health policies?	97.67
5_2_2	Review public health policies at least every two years?	78.33
5_2_3	Advocate for the development of prevention and protection policies?	100
		100
5.3 Community Health Improvement Process 5_3_1	Established a community health improvement process?	100
5_3_2	Developed strategies to address community health objectives?	100
		85.56
5.4 Strategic Planning and Alignment 5_4_1	Each organization in the LPHS conduct a strategic planning process?	100
5_4_2	Each organization in the LPHS review its organizational strategic plan?	56.67
	Local governmental public health entity conduct strategic planning activities?	100

PDFs with Tables – poorly formatted

Shiny Edit Language Models

Shiny Edit Language Models




```
['H', 'e', 'l', '<u>', 'o', ' ', 'h', 'o', 'w', ' ', 'a', 'r', 'e']
['h', 'e', 'l', 'l', 'o', ' ', 'h', 'o', 'w', ' ', 'a', 'r', 'e']
['X', '=', '=', 'I', '=', '=', '=', '=', '=', '=', '=', '=', '=']
```

- Non autoregressive models (alternative to LMs)
- Generative story is as follows:
  - Sample clean GT text (t)
  - Sample an edit vector (condensing all noise) (z)
  - Sample corrupt OCR text given GT and edit vector (x)
- Inference: using VI

$$p(X_{1:N}) = \prod_{n=1}^{N} \sum_{t_n} \int_{z_n} \overset{\text{BERT}}{p(x_n \mid t_n, z_n)}\, p(t_n)\, p(z_n)\, dz_n$$

where, $p(t_n) \sim \mathrm{Dir}(\theta)$ [all vocab]

$p(z_n) \sim \mathrm{VMF}(z)$

VI: $q(t \mid x),\ q(z \mid t, x)$

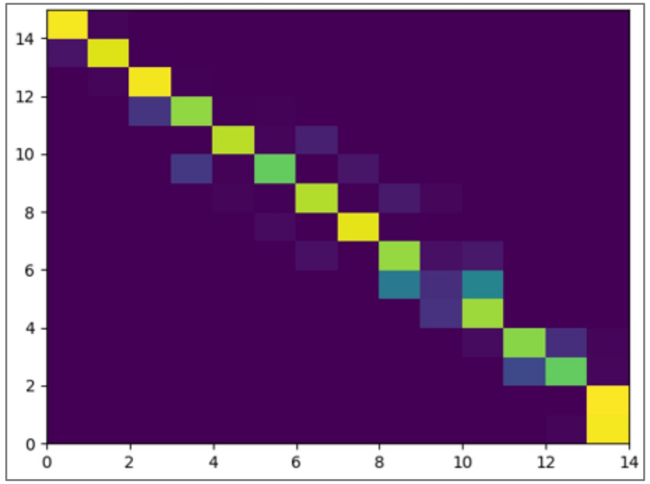
Exhausting all tricks

Diagonal Attention + Coverage

- Given the monotonicity of dependence of sequences in Post-OCR Correction (left to right)
- Word at 4th **position** in output sequence more likely to be dependent on **(3-5) positions** in input sequence
- Stronger dependency than other seq2seq translation scenarios

Copy Mechanism

- Most characters contain no error ; so retain them
- Learn probability **p** such that we retain characters with prob. **p**



BERT Fails

	precision	recall	f1-score	support
BadGT	0.00	0.00	0.00	499
CapsError	0.00	0.00	0.00	99
ExtraContent	0.00	0.00	0.00	456
ExtraSpacing	0.00	0.00	0.00	16
Hyphenation	0.00	0.00	0.00	14
Misrecognition	0.00	0.00	0.00	1404
None	0.93	1.00	0.96	50975
Punctuation	0.00	0.00	0.00	344
RemovedSpacing	0.00	0.00	0.00	831
Shapes	0.00	0.00	0.00	102
UnderScore	0.00	0.00	0.00	56
accuracy			0.93	54796
macro avg	0.08	0.09	0.09	54796
weighted avg	0.87	0.93	0.90	54796

Dataset Statistics

- Processed 832 documents, with 10,000 pages (Average 12 pages per document)
- Resulted in 66644 sentence pairs out of which only 15748 had errors in them -> Most sentences were error free

BERT Fails

Opportunity – Gold JSON files?

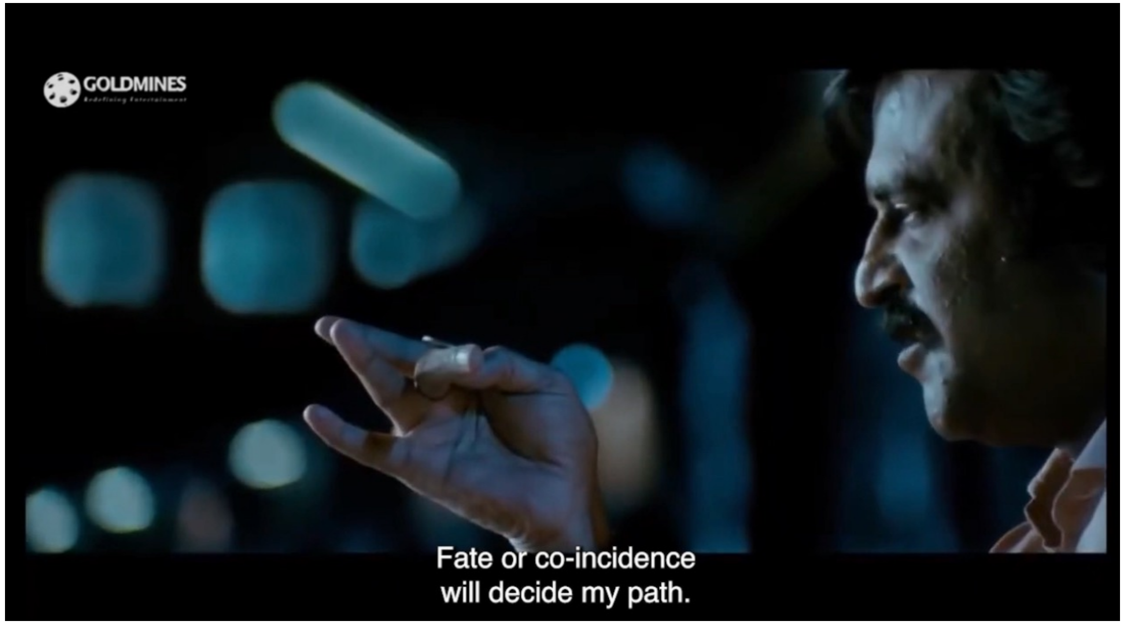
- Get labels of "type" of documents
    - Remove figures
    - Remove tables
    - Retain only paragraphs

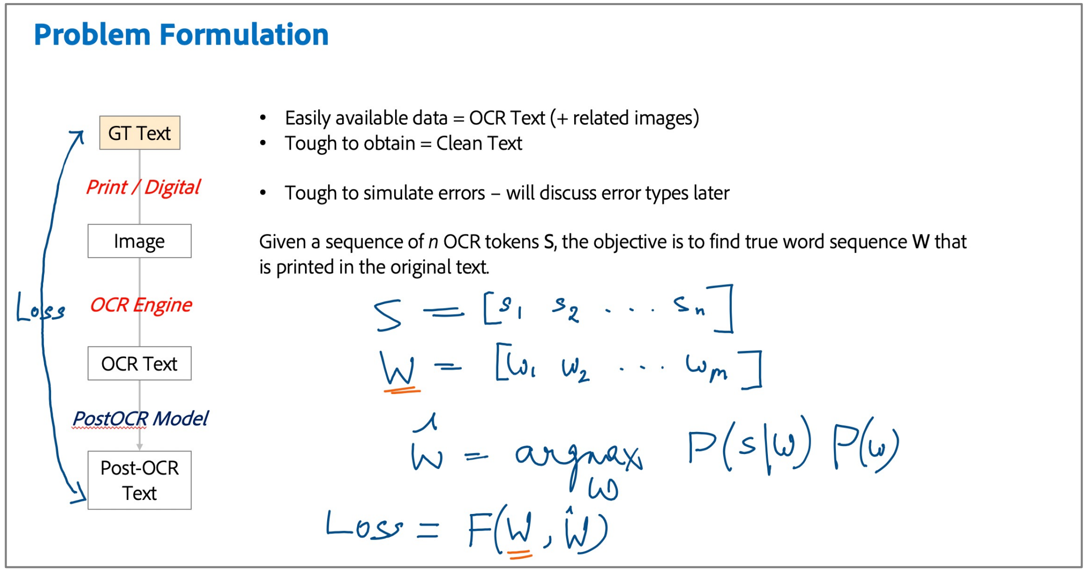
- Get alignments of GT and OCR

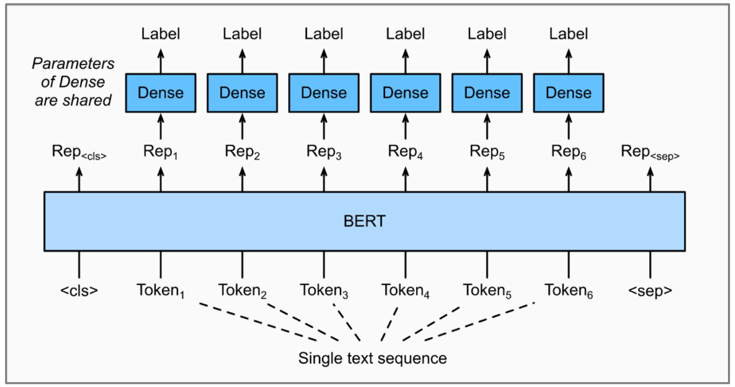
- Get footnotes, sidenotes, etc. separately

- Get confidence scores

Questions?


GOLDMINES
Fate or co-incidence
will decide my path.


Problem Formulation
GT Text
Print / Digital
Image
OCR Engine
OCR Text
PostOCR Model
Post-OCR Text
Loss
• Easily available data = OCR Text (+ related images)
• Tough to obtain = Clean Text
• Tough to simulate errors – will discuss error types later
Given a sequence of n OCR tokens S, the objective is to find true word sequence W that is printed in the original text.
S = [s1 s2 . . . sn]
W = [w1 w2 . . . wm]
Ŵ = argmax P(s|W) P(w)
Loss = F(W, Ŵ)


Label
Dense
Parameters of Dense are shared
Rep<cls>
Rep1
Rep2
Rep3
Rep4
Rep5
Rep6
Rep<sep>
BERT
<cls>
Token1
Token2
Token3
Token4
Token5
Token6
<sep>
Single text sequence

Questions?

BERT

+

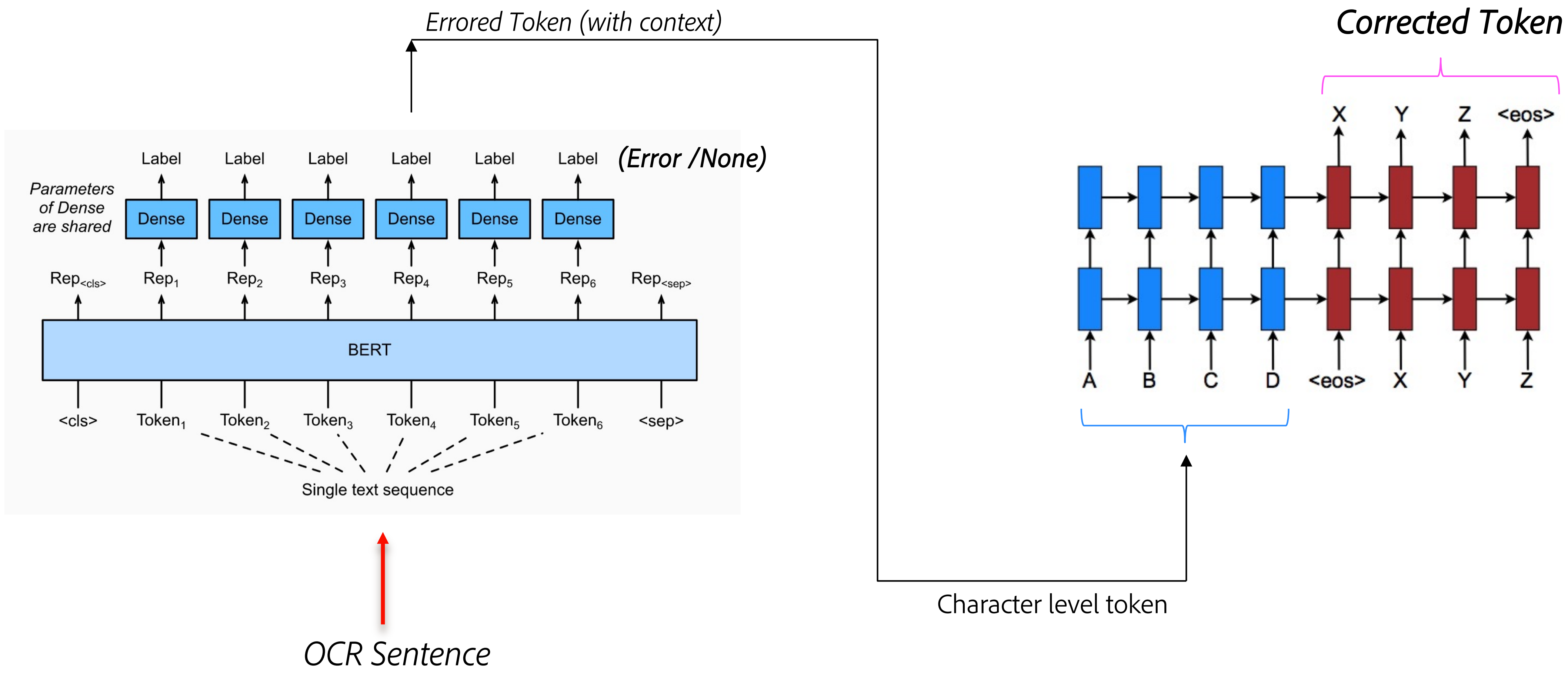
=

PostOCR
BERT

Post-OCR BERT Pipeline

Error Detection

Error Correction



Adobe Scan Dataset

- Filtered Error Statistics

```
({None: 367119,
  'Misrecognition': 2514,
  'ExtraContent': 9002,
  'ContentLoss': 2615,
  'UnderScore': 654,
  'RemovedSpacing': 1008,
  'Punctuation': 738,
  'Hyphenation': 27,
  'CapsError': 37,
  'Shapes': 529,
  'ExtraSpacing': 2},
```

Initial data statistics

```
{None: 105207,
 'Misrecognition': 870,
 'RemovedSpacing': 384,
 'ExtraContent': 65,
 'Shapes': 54,
 'Punctuation': 149,
 'ContentLoss': 112,
 'UnderScore': 27,
 'Hyphenation': 8,
 'CapsError': 15})
```

After filtering and Pre-processing

Adobe Scan Dataset

Training Dataset Details:

- Curated from 400 documents (300 train / 100 test)
- Shortlisted sentences containing errors - 6061 train and 2071 test
- Token Level Error Statistics:

Token Type	Train	Test
None	114232	46136
Error	9030	2947

Evaluation Metrics: Robust Accuracy

Standard Metrics

- Accuracy
- Edit Distance
- Precision, Recall, F1 Score

$$Acc(f) = \mathbb{E}_{(x,y)\sim D}\left(\mathbb{I}\left(f(x) = y\right)\right)$$

Robust Accuracy – Lower bound

- Even if one variation goes wrong, R-Acc reduces

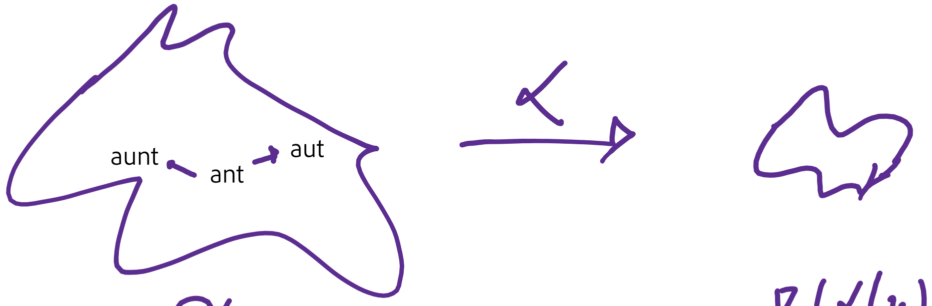
$$R\text{-}Acc(f) = \mathbb{E}_{(x,y)\sim D}\left(\min_{\bar{x} \in B(x)} \mathbb{I}\left(f(\bar{x}) = y\right)\right)$$

Len(S) = 10; Len(w) = 5 (n)

B(x) : Attack surface = Edit Distance – 1

$$O\left(\left({}^{n}C_1 \times 26\right)^{10}\right)$$

$$O\left(5^{10}\right)$$ [Shrinked Attack Surface]



B(X)

B(α(x))

Results - Error Detection

- Detection model will classify the token into Error/None
- Performance:
  - Accuracy: 95%
  - F1 Score: 0.78

	precision	recall	f1-score	support
Error	0.59	0.61	0.60	2947
None	0.97	0.97	0.97	43189
accuracy			0.95	46136
macro avg	0.78	0.79	0.78	46136
weighted avg	0.95	0.95	0.95	46136

Classification Model Performance

Predicted

Actual	None	Error	All
None	41929	1260	43189
Error	1159	1788	2947
All	43088	3028	43136

Confusion Matrix

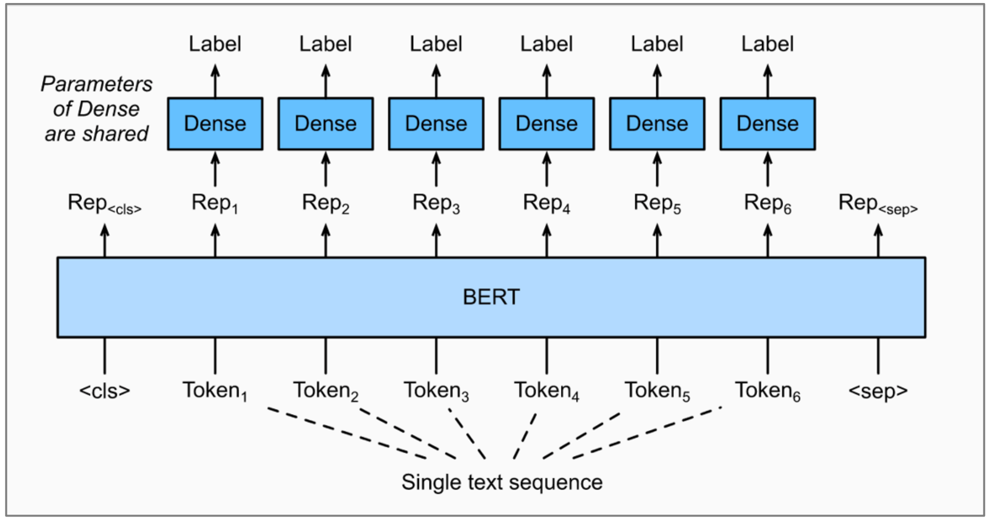
ICDAR Dataset – Detection Results

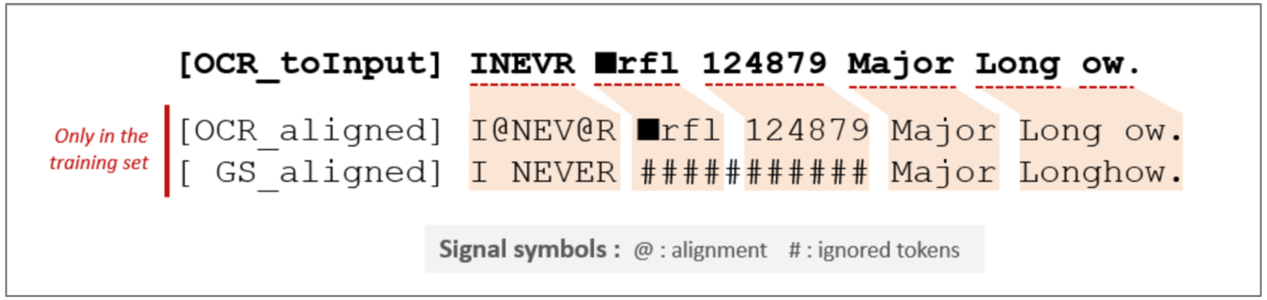
Dataset: ICDAR 2017 Dataset (Public)

- English Monographic sentences
- OCR and GT are aligned at character level
- ~23k sentences (21k/2k split)

Results: Binary Classification task

	precision	recall	f1-score	support
None	0.97	0.98	0.98	58202
error	0.82	0.76	0.79	6254
accuracy			0.96	64456
macro avg	0.90	0.87	0.88	64456
weighted avg	0.96	0.96	0.96	64456





Fully aligned GT and OCR texts
- The only dataset with this feature
- Very difficult and expensive to prepare

Qualitative Examples

OCR : Obviously, your clinicians will need to communicate this to the parents and allow a short but reasonaole time for the parents to be with him pending the extubation.
GT. : Obviously, your clinicians will need to communicate this to the parents and allow a short but reasonable time for the parents to be with him pending the extubation.

OCR: Except where lives can be saved, fire chiefs may now allow buildings to bum rather than risk firefighters' lives.
GT: Except where lives can be saved, fire chiefs may now allow buildings to burn rather than risk firefighters' lives.

OCR: # support to policy making at the national level
GT : # support to policy-making at the national level

--Not detected
--Correctly detected

Takeaways

- Easier to detect and correct Misrecognition errors of high-frequent words
- Tough to capture hyphenation errors
- Not a glaring error; but the PM eval statistics don't capture importance

Qualitative Examples

OCR : 4 The hierarchy of controls is a system widely used irl the petrochemical industry to minimize or elimirlate hazards.

GT : 4 The hierarchy of controls is a system widely used in the petrochemical industry to minimize or eliminate hazards.

OCR: Recognising that, Dr Stephen Playfor, a consultant paediatric intensivist with over 13 years' experience, told me that he considered it wise to move directly to MRI scanning and such was undertaken on J1h February.

GT : Recognising that, Dr Stephen Playfor, a consultant paediatric intensivist with over 13 years' experience, told me that he considered it wise to move directly to MRI scanning and such was undertaken on 7 th February.

--Not detected

--Correctly detected

Takeaways

- Easier to detect and correct Misrecognition errors of high-frequent words
- Tough to capture hyphenation errors
- Not a glaring error; but the PM eval statistics don't capture importance


GOLDMINES
You have insulted me gravely.
It has to be returned.

Qualitative Examples

- Spaces errors are corrected:

OCR : According to the Bahai International Community's United Nations Office, Intelligence Ministry officers, raided the home ofFakhroddin Samini on May 31.

Corrected : According to the Bahai International Community's United Nations Office, Intelligence Ministry officers, raided the home of Fakhroddin Samini on May 31.

OCR : Kitty Ussher was interviewed by Catherine Haddon and Ines on 16th June 2016for the Institute for Government's Ministers Reflect Project

Corrected : Kitty Ussher was interviewed by Catherine Haddon and Ines on 16th June 2016 for the Institute for Government's Ministers Reflect Project

Qualitative Examples

- Spelling Corrections:

OCR : Wet com gluten feed is used extensively in diets for growing and finishing cattle in the Midwest

Corrected : Wet corn gluten feed is used extensively in diets for growing and finishing cattle in the Midwest

OCR : Part Ill: shaping the duty to accomodate

Corrected : Part III: shaping the duty to accomodate

Spelling and Spaces:

OCR : Department of Probation, as well as the Mayor's Office oflmmigration Affairs

Corrected : Department of Probation, as well as the Mayor's Office of Immigration Affairs

Results – Error Correction

Total Test Results (Token Level)

- Total errored tokens correctly detected = 1788
- Tokens Corrected to GT = 177
- Accuracy = 9.89 %

Levenstein Distance: (lower score is better)

- Before Correction – 4.77
- After Correction – 3.22

Limitation 1: Bad Ground Truth Text

Bad Ground Truth Text

Debra.Wallace@csun.edu ----> Debra. Wallace@ csun. edu
1PM ----> IPM
always, ----> alwa s,
interplay ----> interpla
immunity, ----> immunit ,
rely ----> rel
says ----> sa s
why ----> wh
clearly ----> clearl
by ----> b
may ----> ma

OCR Text. ---> GT Text

Despite OCR being somewhat correct,
GT misses a “y” – coule be an artifact of the dataset

Bad Ground Truth Text

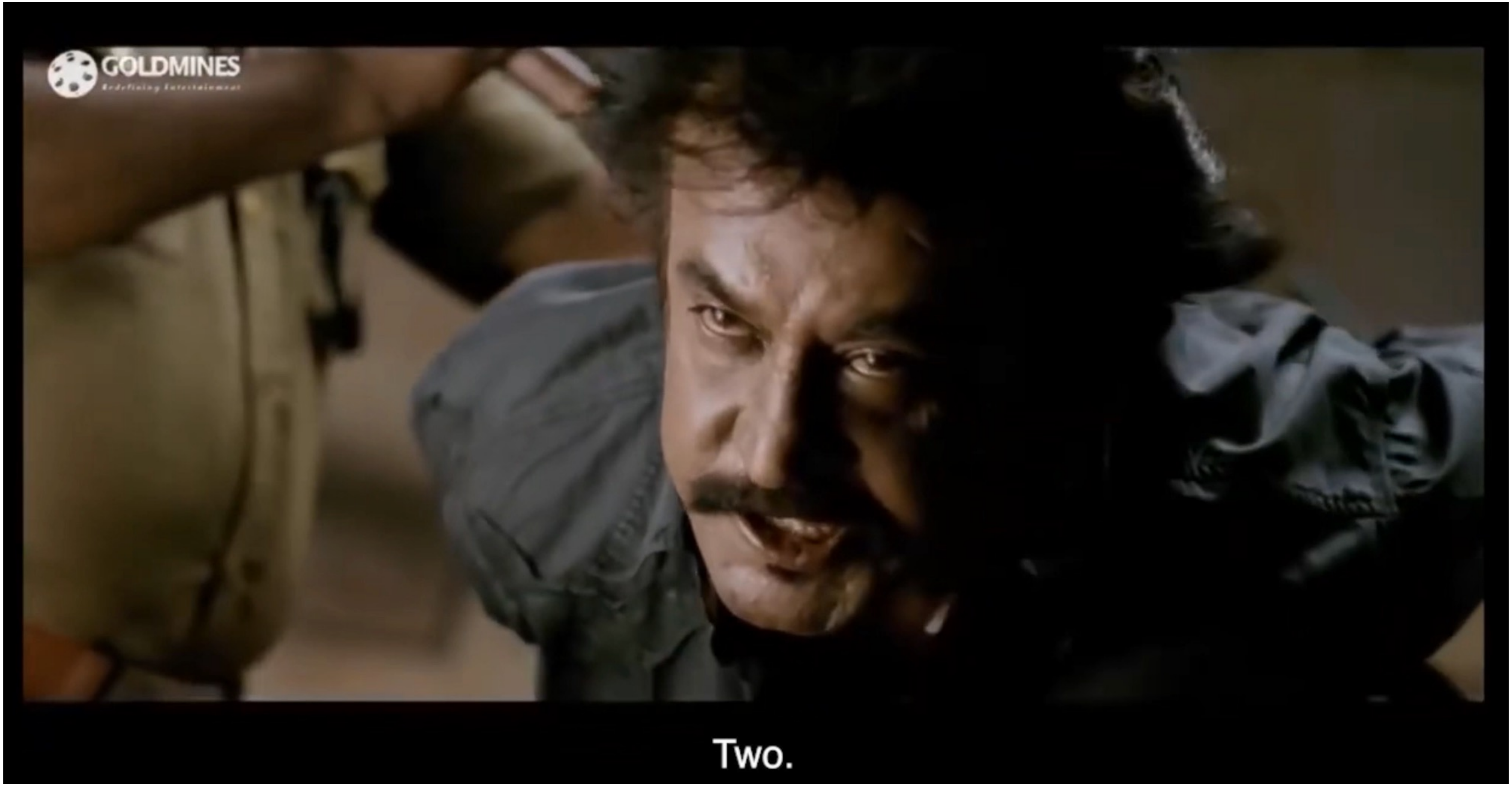
OCR Text [url] with bounding boxes

```
http://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network%x00AD;  0.256078 0.259394 0.819608 0.273939
MLN/MLNProducts/downloads/MedicareRemit  0.216078 0.276667 0.565490 0.288182
```

Ground Truth text - all split up at hyphens

```
http:// 0.254902 0.253727 0.299536 0.275182
www. 0.299534 0.253727 0.341528 0.275182
cms. 0.341526 0.253727 0.378166 0.275182
gov/ 0.378164 0.253727 0.411217 0.275182
Outreach- 0.411215 0.253727 0.487139 0.275182
and- 0.487137 0.253727 0.521072 0.275182
Education/ 0.521070 0.253727 0.603294 0.275182
Medicare- 0.603294 0.253727 0.678352 0.275182
Learning- 0.678352 0.253727 0.751607 0.275182
Network 0.751607 0.253727 0.815038 0.275182
MLN/ 0.214706 0.271076 0.253994 0.292530
MLNProducts/ 0.253992 0.271076 0.362949 0.292530
downloads/ 0.362947 0.271076 0.451365 0.292530
MedicareRemit_ 0.451361 0.271076 0.574623 0.292530
0408. 0.574619 0.271076 0.614838 0.292530
pdf. 0.614835 0.271076 0.644373 0.292530
```

Limitation 2: Bad Alignment



Bad Alignment


```
  matchId: 2
  pctIOU: 88
  parentIOU: 83
▸ elementIds: [] 2 items
▸ overlaps: {} 2 keys
  tagName: "LI"
▾ text: {} 2 keys
    gold: "•Obtaining of additional/bogus load tickets"
    test: "• ing of additional/bogus load tickets"
▸ location: {} 2 keys
▾ differences: [] 3 items
    0: "mediumDiffIOU"
    1: "textContent"
    2: "layout"
```


Sometimes the Gold JSON files have wrong alignment too – "obtain" is there in the previous text dictionary


```
  matchId: 1
  pctIOU: 72
  parentIOU: 10
▸ elementIds: [] 2 items
▸ overlaps: {} 2 keys
▸ tagName: {} 2 keys
▾ text: {} 2 keys
    gold: "□Fraud Conspiracies"
    test: "• I raud Cons roe s"
▸ location: {} 2 keys
▾ differences: [] 5 items
    0: "largeDiffIOU"
    1: "tagName"
    2: "textContent"
    3: "grouping"
    4: "layout"
```

Limitation 3: Boundary Errors

Boundary Errors

```
INTRODUCTION
Although the l i terature dealing with formal and natural
languages abounds with theoretical arguments of worst-
case performance by various parsing strategies \[e.g.,
Griffiths & Petrick, 1965; Aho & Ullman, 1972; Graham,
Harrison & Ruzzo, Ig80\], there is l i t t le discussion of
comparative performance based on actual practice in
understanding natural language. Yet important practical
considerations do arise when writing programs to under-
stand one aspect or another of natural language utteran-
ces. Where, for example, a theorist wi l l characterize a
parsing strategy according to its space and/or time
requirements in attempting to analyze the worst possible
input acc3rding to ~n arbitrary grammar st r ic t ly limited
in expressive power, the researcher studying Natural
Language Processing can be just i f ied in concerning
himself more with issues of practical performance in
parsing sentences encountered in language as humans
Actually use i t using a grammar expressed in a form
corve~ie: to the human linguist who is writing i t .
```

```
>>> from Bio.pairwise2 import format_alignment
>>> print(format_alignment(*alignments[0]))
ACCGT
| ||
A-CG-
  Score=3
```

- Segmentation boundary errors are very difficult to correct: They may seem “non-word” ; but can quickly turn into ”correct word” by deleting / inserting a whitespace at appropriate position

- GT: LITTLE; OCR: L_I_T_TL_E

- How to map each character to its corresponding correction? Teach model to predict “noop” character @
- (L_I_T_TL_E) → [(L@I@T@TL@E) == (LITTLE)]

- Alignment
    - RETAS Scheme: Recursive Text Alignment
        - Finds unique words common to both texts and uses as anchor points
    - Needleman Wunsch Algorithm (BioPython)

- **Hurts both training and evaluation for longer texts**
    - Need ICDAR-like span labelling

Limitations

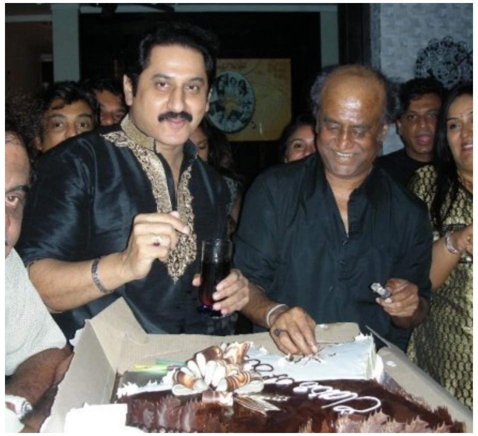
1. Bad Ground Truth Text

2. Bad Alignment

3. Boundary Errors – tough to prepare train / eval

=> Shortlisted files via Gold JSON alignment have very few relevant errors; leading to poor eval scores

Partners in Crime?
It could be the case that both OCR and PostOCR suffer from similar pathologies



Adobe

Can BERT make a comeback? Genalog + Vistext Embeddings