European Commission

HUN-REN Alfréd Rényi Institute of Mathematics
Artificial Intelligence Group

ÁBTL Állambiztonsági Szolgálatok Történeti Levéltára

MNL Magyar Nemzeti Levéltár

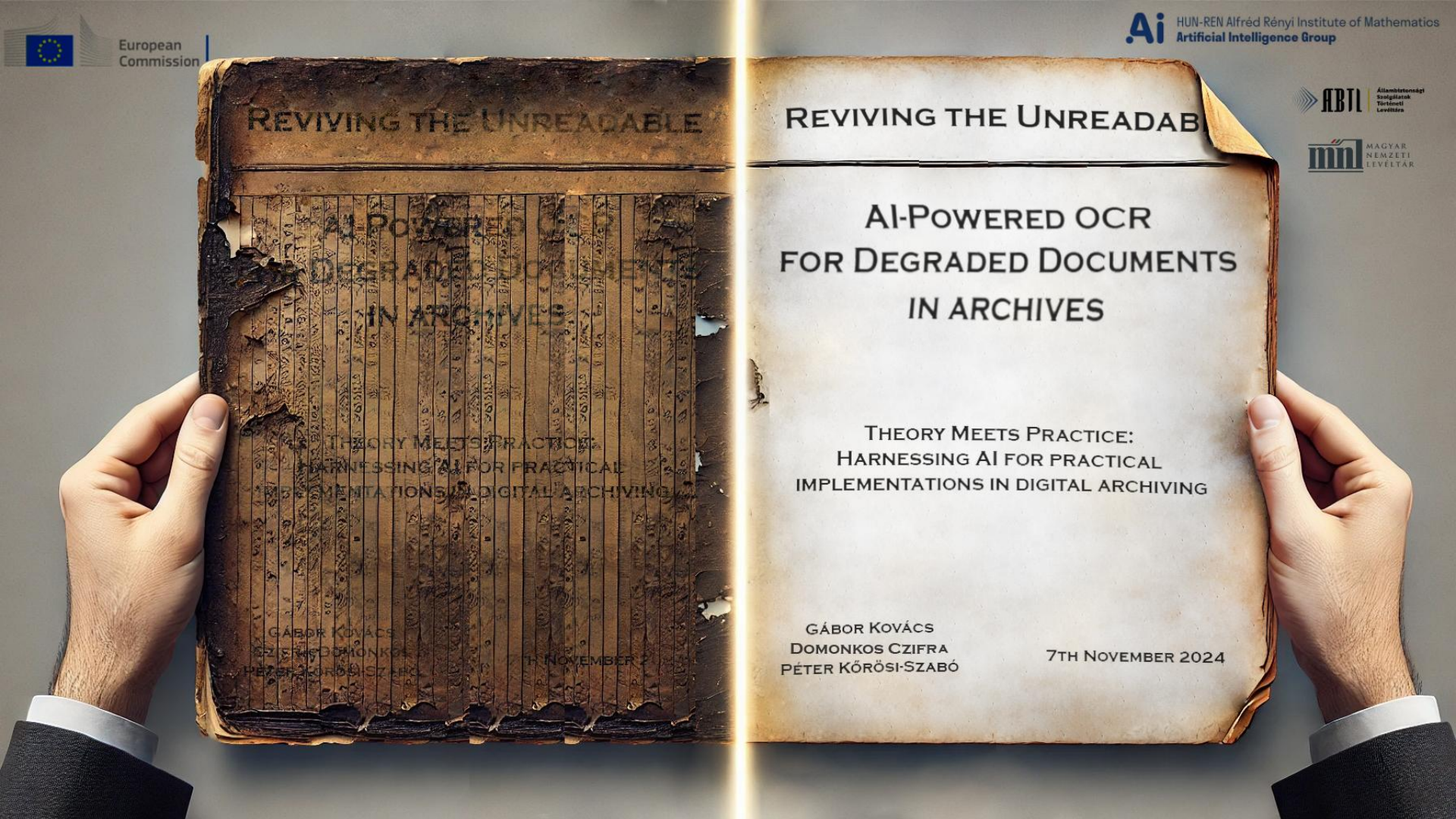
Reviving the Unreadable

AI-Powered OCR for Degraded Documents in Archives

Theory Meets Practice: Harnessing AI for practical implementations in digital archiving

Gábor Kovács
Domonkos Czifra
Péter Kőrösi-Szabó

7th November 2024


European Commission
HUN-REN Alfréd Rényi Institute of Mathematics
Artificial Intelligence Group
ÁBTL
Állambiztonsági Szolgálatok Történeti Levéltára
MNL
Magyar Nemzeti Levéltár
Reviving the Unreadable
AI-Powered OCR for Degraded Documents in Archives
Theory Meets Practice: Harnessing AI for practical implementations in digital archiving
Gábor Kovács
Domonkos Czifra
Péter Kőrösi-Szabó
7th November 2024

Rényi AI: Introduction

Alfréd Rényi Institute of Mathematics:

- Artificial Intelligence group

Theoretical Research Areas:

- Computer Vision
- Learning Theory

Applied (Industrial) Projects:

- Time Series Prediction:
  - Energy sector
  - Medical sector
- **Natural Language Processing (NLP) and Document AI**
  - **Hungarian Archives:**
    - Historical Archives of the Hungarian State Security
    - National Archives of Hungary

Applied Projects with Archives

Main goal:

- Integrate AI solutions into archival processes

AI solutions:

- Latest advances in AI technology:
  - Large Language Models (LLMs)

Applications:

- Retrieval Augmented Generation (RAG)
- Sensitive Data Removal
  - GDPR, inquiry, research
- Named Entity Recognition
  - Database, Knowledge Graph

Tasks & Challenges

Tasks and Challenges

Archive specific challenges:

- **Data format**:
  - We have: scanned document images
  - We need: text-based data
- **Data quality**:
  - highly degraded documents

Tasks and Challenges

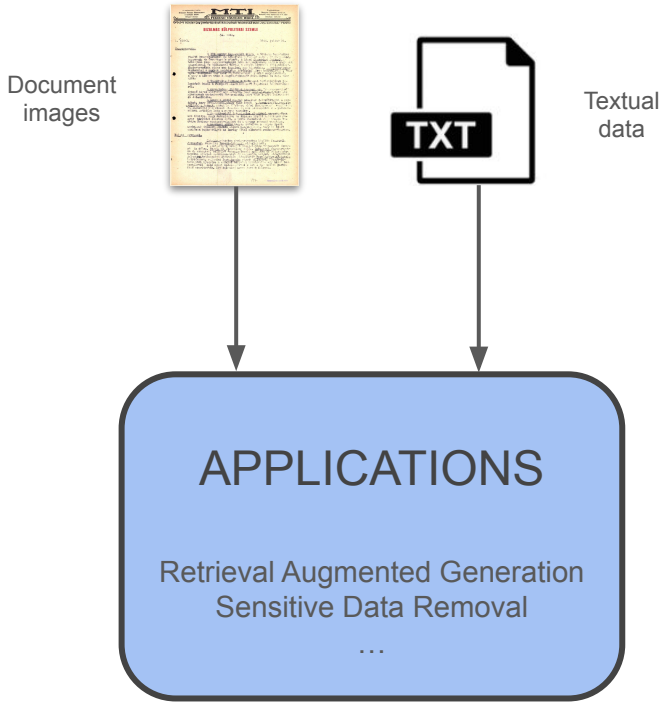
Archive specific challenges:

- **Data format**:
  - We have: scanned document images
  - We need: text-based data
- **Data quality**:
  - highly degraded documents

Tasks and Challenges

Archive specific challenges:

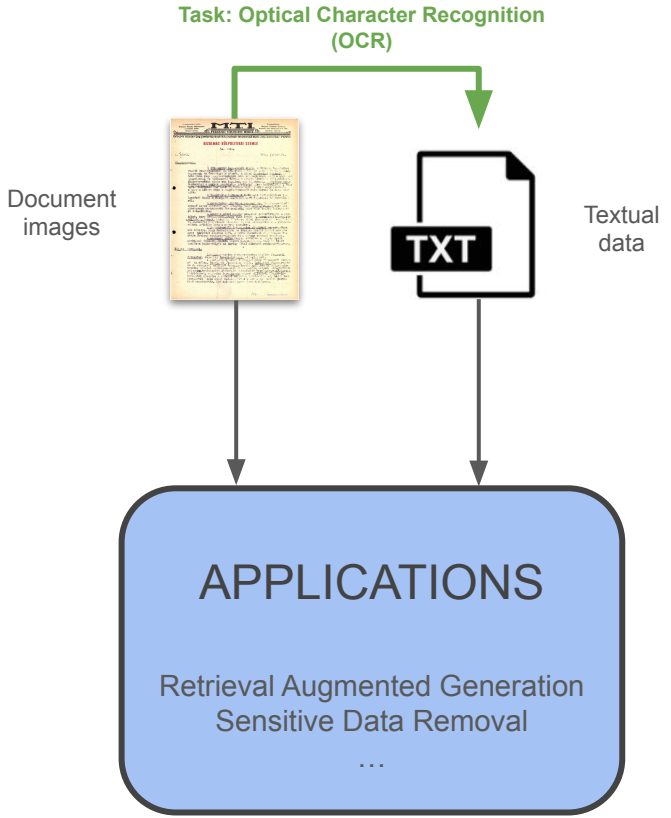
- **Data format**:
  - We have: scanned document images
  - We need: text-based data
- **Data quality**:
  - highly degraded documents



Tasks and Challenges

Archive specific challenges:

- **Data format**:
  - We have: scanned document images
  - We need: text-based data
- **Data quality**:
  - highly degraded documents



Tasks and Challenges

Archive specific challenges:

- **Data format**:
  - We have: scanned document images
  - We need: text-based data
- **Data quality**:
  - highly degraded documents

European Commission

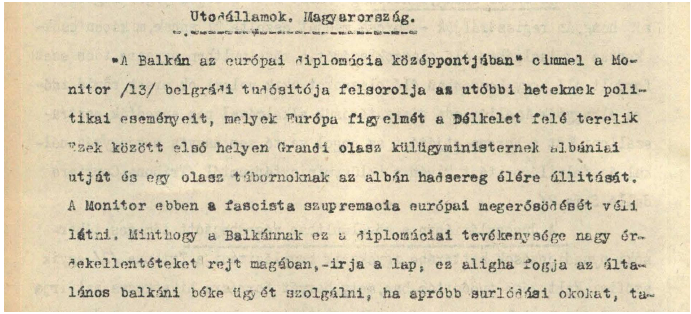
HUN-REN Alfréd Rényi Institute of Mathematics
Artificial Intelligence Group

Tasks and Challenges

Archive specific challenges:

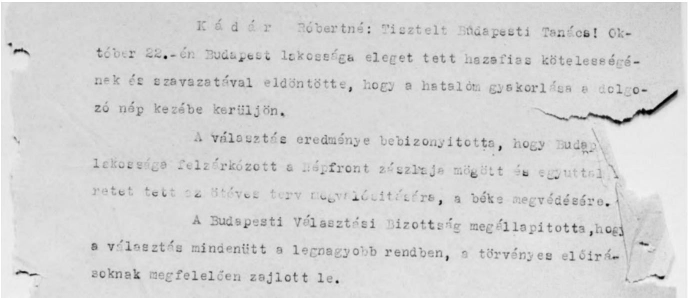
- **Data format**:
    - We have: scanned document images
    - We need: text-based data
- **Data quality**:
    - highly degraded documents

Good


Utoódállamok. Magyarország.

"A Balkán az európai diplomácia középpontjában" cimmel a Monitor /13/ belgrádi tudósitója felsorolja az utóbbi heteknek politikai eseményeit, melyek Európa figyelmét a Délkelet felé terelik Ezek között első helyen Grandi olasz külügyminiszternek albániai utját és egy olasz tábornoknak az albán hadsereg élére állitását. A Monitor ebben a fascista szupremacia európai megerősödését véli látni. Minthogy a Balkánnak ez a diplomáciai tevékenysége nagy érdekellentéteket rejt magában,-irja a lap, ez aligha fogja az általános balkáni béke ügyét szolgálni, ha apróbb surlódási okokat, ta-

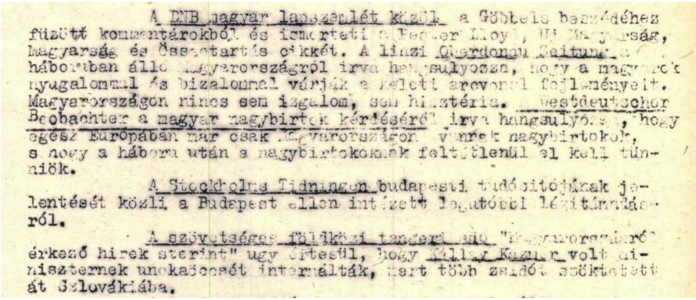
Moderate


Kádár Róbertné: Tisztelt Budapesti Tanács! Október 22.-én Budapest lakossága eleget tett hazafias kötelességének és szavazatával eldöntötte, hogy a hatalom gyakorlása a dolgozó nép kezébe kerüljön.

A választás eredménye bebizonyitotta, hogy Budap lakossága felzárkózott a Népfront zászlaja mögött és egyuttal retet tett az ötéves terv megvalósitására, a béke megvédésére.

A Budapesti Választási Bizottság megállapitotta, hogy a választás mindenütt a legnagyobb rendben, a törvényes előirásoknak megfelelően zajlott le.

Poor


A DNB magyar lapszemlét közli a Göbbels beszédéhez fűzött kommentárokból és ismerteti a Pester Lloyd, Uj Magyarság, Magyarság és Összetartás cikkét. A linzi Oberdonau Zeitung a háborúban álló Magyarországról irva hangsulyozza, hogy a magyarok nyugalommal és bizalommal várják a keleti arcvonal fejleményeit. Magyarországon nincs sem izgalom, sem hisztéria. A Westdeutscher Beobachter a magyar nagybirtok kérdéséről irva hangsulyozza, hogy egész Európában már csak Magyarországon vannak nagybirtokok, s hogy a háboru után a nagybirtokoknak feltétlenül el kell tünniök.

A Stockholms Tidningen budapesti tudósitójának jelentését közli a Budapest ellen intézett legutóbbi légitámadásról.

A szövetséges földközi tengeri ... "Magyarországról érkező hirek szerint" ugy értesül, hogy Kállay Kazimir volt miniszternek unokaöccsét internálták, mert több zsidót szöktetett át Szlovákiába.

D e m o

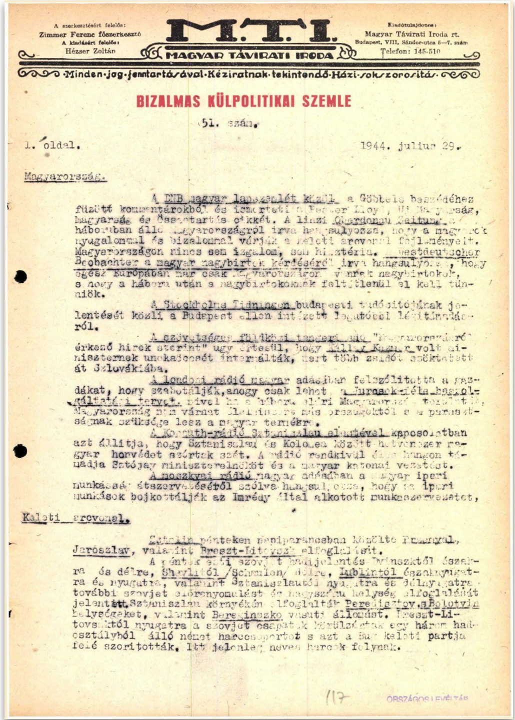
How SOTA methods struggle to extract text

European Commission

HUN-REN Alfréd Rényi Institute of Mathematics
Artificial Intelligence Group

Demo: SOTA off-the-shelf OCR systems

Original Document



A szerkesztésért felelős: Zimmer Ferenc főszerkesztő
A kiadásért felelős: Hézser Zoltán

M.T.I.
MAGYAR TÁVIRATI IRODA

Kiadótulajdonos: Magyar Távirati Iroda rt.
Budapest, VIII, Sándor-utca 5—7. szám
Telefon: 145-510

Minden jog fenntartásával · Kéziratnak tekintendő · Házi sokszorosítás

BIZALMAS KÜLPOLITIKAI SZEMLE

51. szám.

1. oldal. 1944. julius 29.

Magyarország.

A DNB magyar lapszemlét közöl a Göbbels beszédéhez fűzött kommentárokból és ismerteti a Pester Lloyd, Uj Magyarság, Magyarság és Összetartás cikkét. A linzi Oberdonau Zeitung a háborúban álló Magyarországról irva hangsulyozza, hogy a magyarok nyugalommal és bizalommal várják a keleti arcvonal fejleményeit. Magyarországon nincs sem izgalom, sem hisztéria. A Westdeutscher Beobachter a magyar nagybirtok kérdéséről irva hangsulyozza, hogy egész Európában már csak Magyarországon vannak nagybirtokok, s hogy a háboru után a nagybirtokoknak feltétlenül el kell tűnniök.

A Stockholms Tidningen budapesti tudósitójának jelentését közli a Budapest ellen intézett legutóbbi légitámadásról.

A szövetséges földközi tengeri adó "Magyarországról érkező hirek szerint" ugy értesül, hogy Kállay Kazmér volt miniszternek unokaöccsét internálták, mert több zsidót szöktetett át Szlovákiába.

A londoni rádió magyar adásában felszólitotta a gazdákat, hogy szabotálják, amogy csak lehet, a Jurcsek-féle beszolgáltatási tervet, mivel ha a háboru eléri Magyarország területét, Magyarország nem várhat élelmiszert más országoktól s a parasztságnak szüksége lesz a magyar termésre.

A Kossuth-rádió Sztaniszlau elestével kapcsolatban azt állitja, hogy Sztaniszlau és Kolomea között hatvanezer magyar honvédet szórtak szét. A rádió rendkivül éles hangon támadja Sztójay miniszterelnököt és a magyar katonai vezetést.

A moszkvai rádió magyar adásában a magyar ipari munkásság átszervezéséről szólva hangsulyozza, hogy az ipari munkások bojkottálják az Imrédy által alkotott munkaszervezetet,

Keleti arcvonal.

Sztalin pénteken napiparancsban közölte Przemysl, Jaroszlav, valamint Breszt-Litovszk elfoglalását.

A pénteki esti szovjet hadijelentés Dvinszktől északra és délre, Shavlitól /Schaulen/ délre, Lublintól északnyugatra és nyugatra, valamint Sztaniszlautól nyugatra és délnyugatra további szovjet előrenyomulást és nagyszámu helység elfoglalását jelentett. Sztaniszlau környékén elfoglalták Peredisztov, Bolotvin helységeket, valamint Bereginszko vasuti állomást. Breszt-Litovszktól nyugatra a szovjet csapatok körülzártak egy három hadosztályból álló német harccsoportot s azt a Bug keleti partja felé szoritották. Itt jelenleg heves harcok folynak.

117

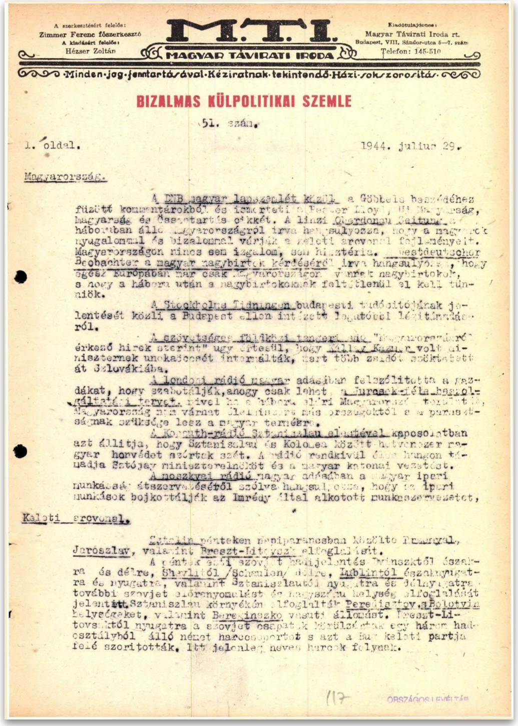
European Commission

HUN-REN Alfréd Rényi Institute of Mathematics
Artificial Intelligence Group

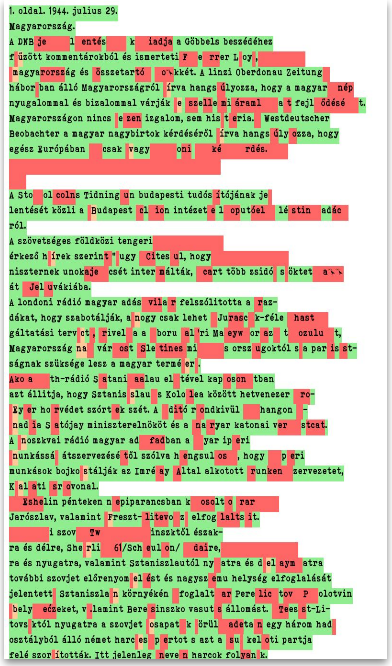
Demo: SOTA off-the-shelf OCR systems

- Correct
- Small error
- Large error

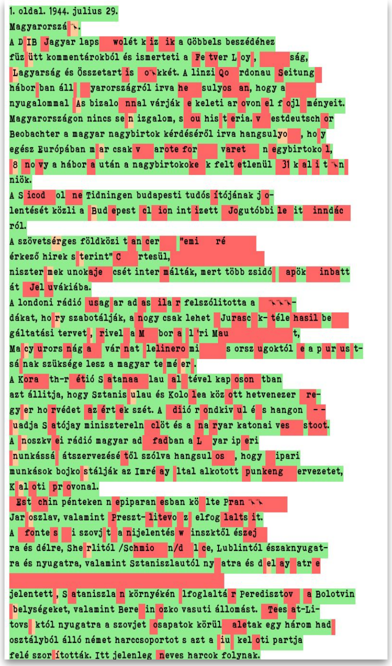
Original Document



Gemini



Google Lens



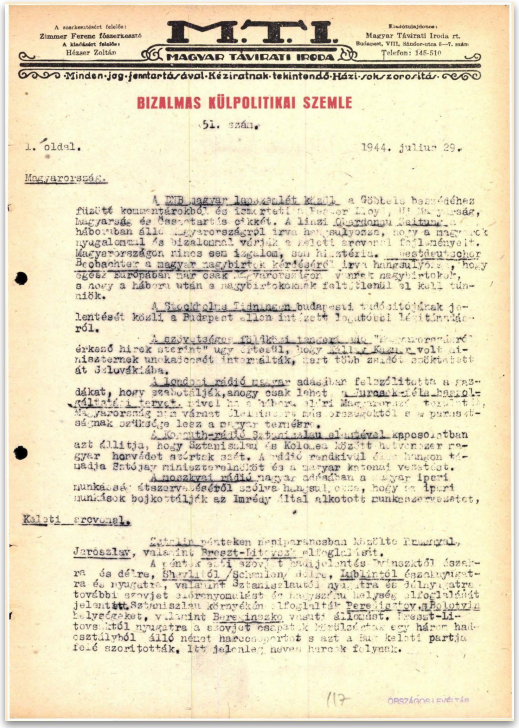
Demo: SOTA off-the-shelf OCR systems

- Correct
- Small error
- Large error

CER: Character Error Rate

WER: Word Error Rate

Original Document



M.T.I.

MAGYAR TÁVIRATI IRODA

Minden jog fenntartóval · Kéziratnak tekintendő · Házi zokzorosító

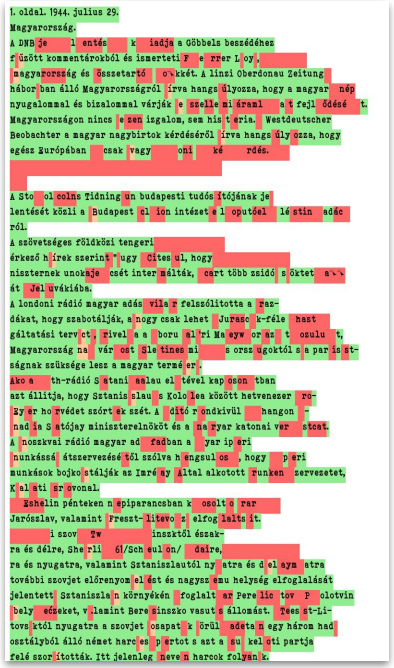
BIZALMAS KÜLPOLITIKAI SZEMLE

51. szám.

1. oldal. 1944. július 29.

Magyarország.

Gemini

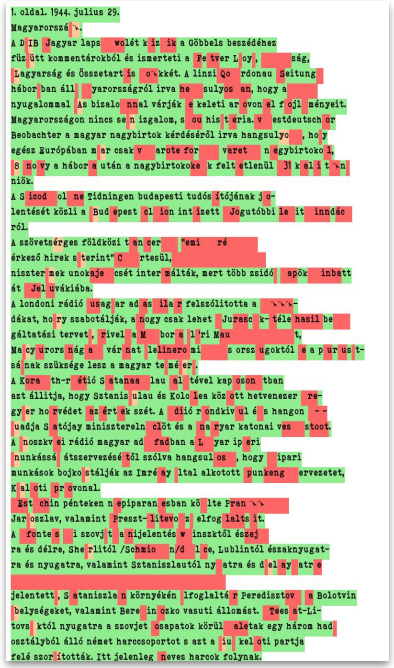


1. oldal. 1944. július 29.

Magyarország.

CER: 25.1%
WER: 56.9%

Google Lens



1. oldal. 1944. július 29.

Magyarország.

CER: 22.9%
WER: 60.5%

Error Analysis

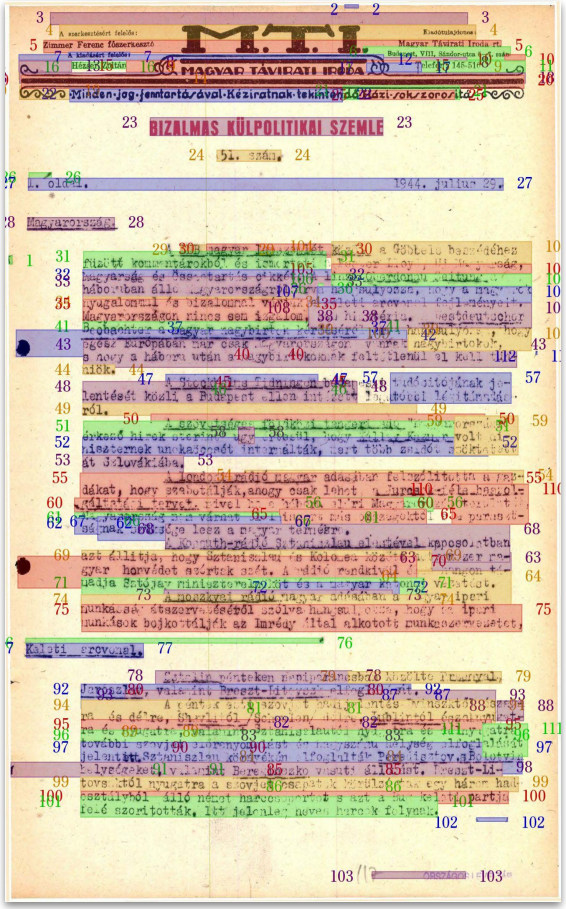
Reasons behind the poor performance:

- **Text Detection**:
  - challenging to detect text lines
  - some part of the text is already lost
- **Text Recognition**:
  - challenging to read out text

Error Analysis

Reasons behind the poor performance:

- **Text Detection**:
  - challenging to detect text lines
  - some part of the text is already lost
- **Text Recognition**:
  - challenging to read out text



missing text

false positive

overlapping boxes

inconsistent size

Error Analysis

Reasons behind the poor performance:

- **Text Detection**:
  - challenging to detect text lines
  - some part of the text is already lost

(GT) A szövetséges földközi tengeri adó “Magyarországról

(OCR) A szövetsérges földközi t an cer “emi ré

- **Text Recognition**:
  - challenging to read out text

(GT) gyar honvédet szórtak szét. A rádió rendkivül éles hangon tá-

(OCR) gy er ho rvédet az ért ek szét. A diió r ondkiv ul é s hangon - -

Our Solution

Approaches

Enhancing OCR systems directly:

- Data perspective:
  - archival data domain
- Architecture perspective:
  - larger models
  - change model type
    - e.g. Transformers

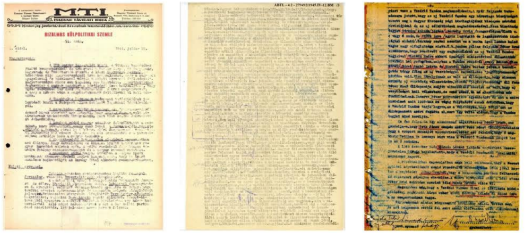
Preprocessing for OCR:

- Document Image Enhancement (DIE)
  - improving the quality of the input image

Approaches

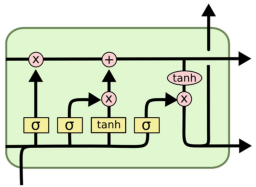
Enhancing OCR systems directly:

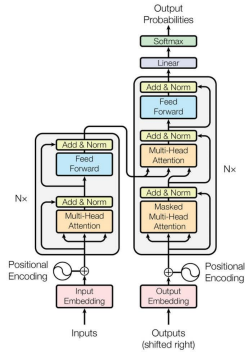
- Data perspective:
  - archival data domain
- Architecture perspective:
  - larger models
  - change model type
    - e.g. Transformers



Preprocessing for OCR:

- Document Image Enhancement (DIE)
  - improving the quality of the input image





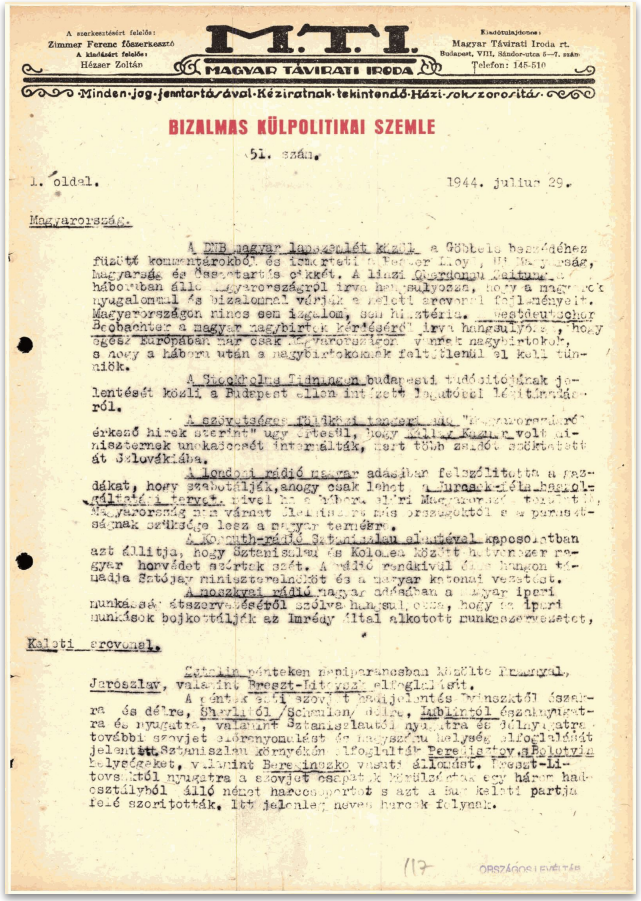
Approaches

Enhancing OCR systems directly:

- Data perspective:
  - archival data domain
- Architecture perspective:
  - larger models
  - change model type
    - e.g. Transformers

Preprocessing for OCR:

- Document Image Enhancement (DIE)
  - improving the quality of the input image



M.T.I.

MAGYAR TÁVIRATI IRODA

Minden jog fenntartásával. Kéziratnak tekintendő. Házi sokszorosító.

BIZALMAS KÜLPOLITIKAI SZEMLE

51. szám.

1. oldal. 1944. julius 29.

Magyarország.

Kalati arcvonal.

Solution

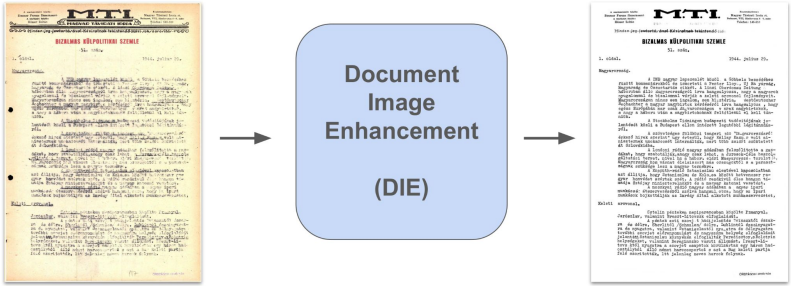
Document Image Enhancement (DIE):

- **Cleaning document images**:
  - removing unnecessary noise while retaining text
- **Model architecture**:
  - convolutional based U-Net
    - U-Net: image transformation, removing noise
    - Convolution operation: filtering image
    - Skip connection: better training properties

Solution

Document Image Enhancement (DIE):

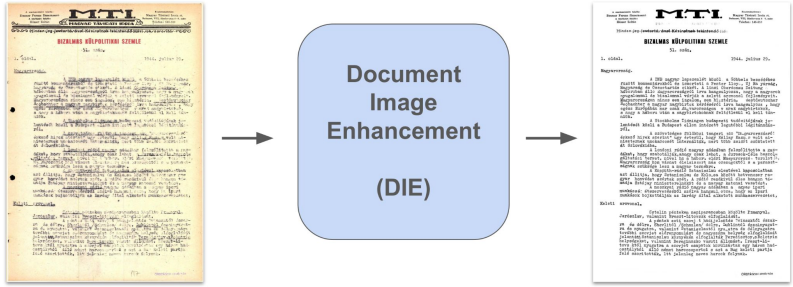
- **Cleaning document images**:
  - removing unnecessary noise while retaining text
- **Model architecture**:
  - convolutional based U-Net
    - U-Net: image transformation, removing noise
    - Convolution operation: filtering image
    - Skip connection: better training properties

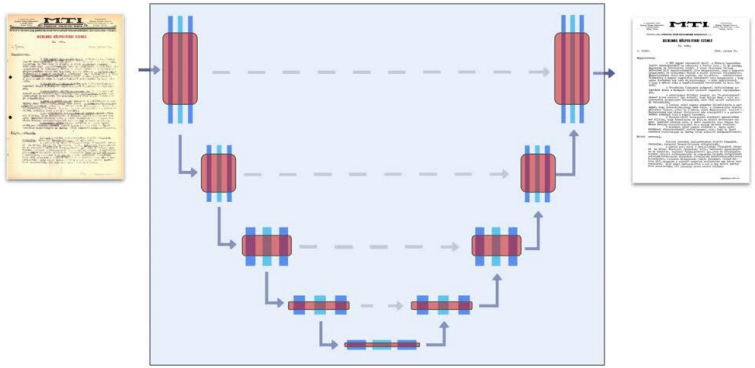


Solution

Document Image Enhancement (DIE):

- **Cleaning document images**:
  - removing unnecessary noise while retaining text
- **Model architecture**:
  - convolutional based U-Net
    - U-Net: image transformation, removing noise
    - Convolution operation: filtering image
    - Skip connection: better training properties





D e m o

Demo: Document Image Enhancement (DIE)

Data

Data

Machine Learning perspective:

- Machine Learning (ML) paradigm:
  - solving the task by learning from examples / data
- ML model needs data:
  - Text Detection: bounding boxes around the text lines
  - Text Recognition: text in the document
  - Document Image Enhancement: clean version (black text on white background)

Supervised Learning:

- Human labeling:
  - Text Detection: cumbersome
  - Text Recognition: cumbersome
  - Document Image Enhancement: extreme hard
- Synthetic generation:
  - Text Detection: easy
  - Text Recognition: easy
  - Document Image Enhancement: moderate

Data

Machine Learning perspective:

- Machine Learning (ML) paradigm:
  - solving the task by learning from examples / data
- ML model needs data:
  - Text Detection: bounding boxes around the text lines
  - Text Recognition: text in the document
  - Document Image Enhancement: clean version (black text on white background)

Supervised Learning:

- Human labeling:
  - Text Detection: cumbersome
  - Text Recognition: cumbersome
  - Document Image Enhancement: extreme hard
- Synthetic generation:
  - Text Detection: easy
  - Text Recognition: easy
  - Document Image Enhancement: moderate

Data

Machine Learning perspective:

- Machine Learning (ML) paradigm:
  - solving the task by learning from examples / data
- ML model needs data:
  - Text Detection: bounding boxes around the text lines
  - Text Recognition: text in the document
  - Document Image Enhancement: clean version (black text on white background)

Supervised Learning:

- Human labeling:
  - Text Detection: cumbersome
  - Text Recognition: cumbersome
  - Document Image Enhancement: extreme hard
- Synthetic generation:
  - Text Detection: easy
  - Text Recognition: easy
  - Document Image Enhancement: moderate

Synthetic Data Generation framework

Text:

- language
- layout
- fonts

Background:

- paper type
- overbleed

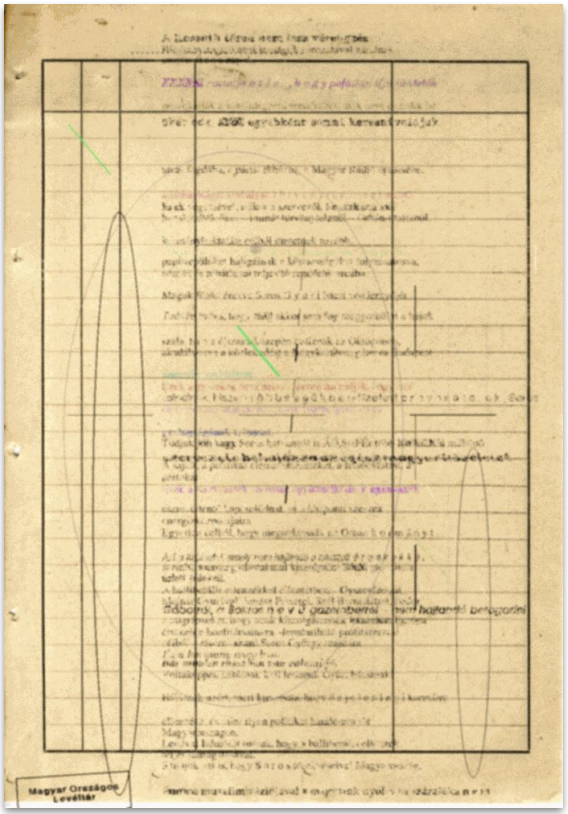
Perturbations:

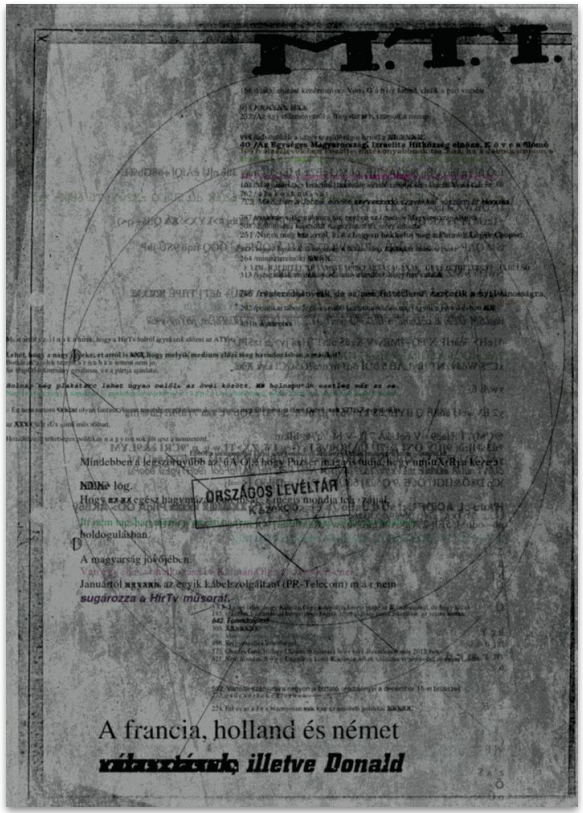
- 20 noise types
- aged document

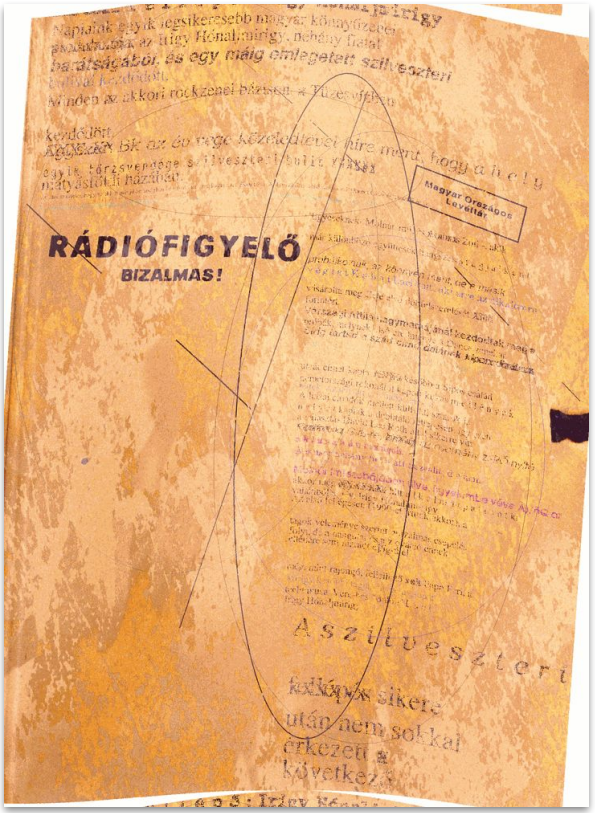
Geometric:

- rotation
- warp

Demo: Synthetic Data Generation







How to customize the framework?

Corpus:

- different kinds of languages

Text types:

- different kinds of fonts

Background images:

- paper types

Noise types:

- stamps, scribbles, textures, …

How to customize the framework?

Corpus:

- different kinds of languages

Text types:

- different kinds of fonts

Background images:

- paper types

Noise types:

- stamps, scribbles, textures, …

How to customize the framework?

Corpus:

- different kinds of languages

Text types:

- different kinds of fonts

Abadi MT Condensed Light
Albertus Extra Bold
Albertus Medium
Antique Olive
Arial
Arial Black
Arial MT

Background images:

- paper types

Noise types:

- stamps, scribbles, textures, …

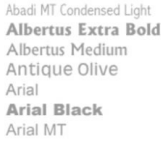
How to customize the framework?

Corpus:

- different kinds of languages

Text types:

- different kinds of fonts



Background images:

- paper types

Noise types:

- stamps, scribbles, textures, …

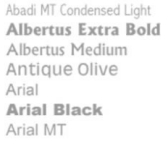
How to customize the framework?

Corpus:

- different kinds of languages

Text types:

- different kinds of fonts

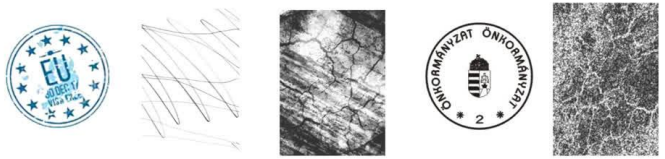


Background images:

- paper types

Noise types:

- stamps, scribbles, textures, …

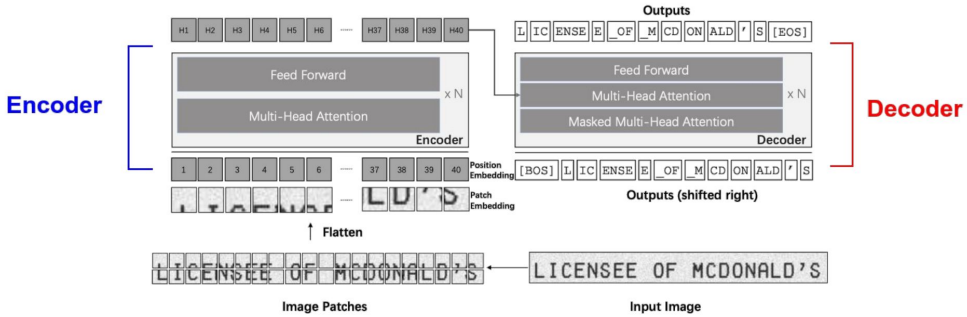


Our OCR

Our OCR model

Training our custom OCR model:

- TrOCR model
  - Transformer-based model
- Trained on:
  - Hungarian language
  - Degraded archival documents
    - Synthetic Data Generator
    - Human labeled data

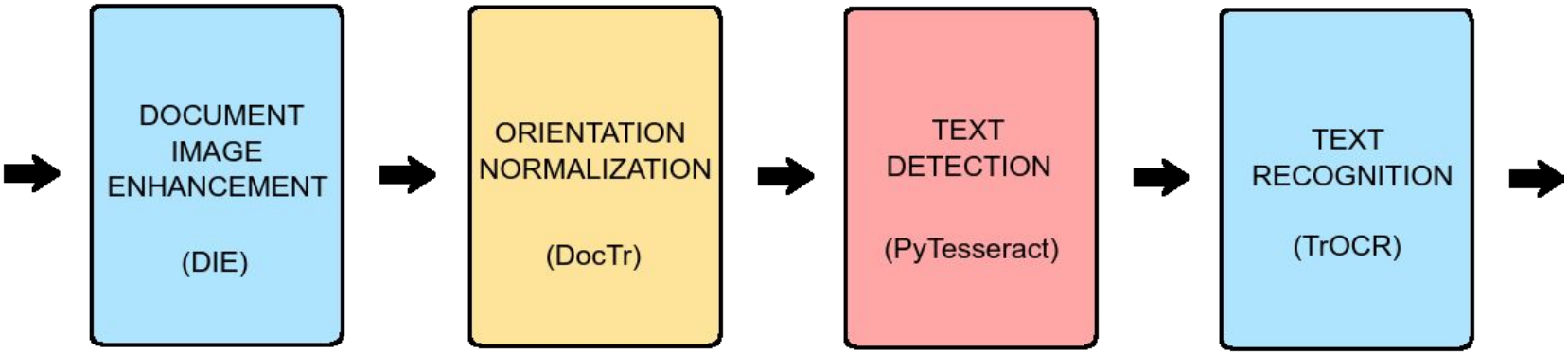


European Commission

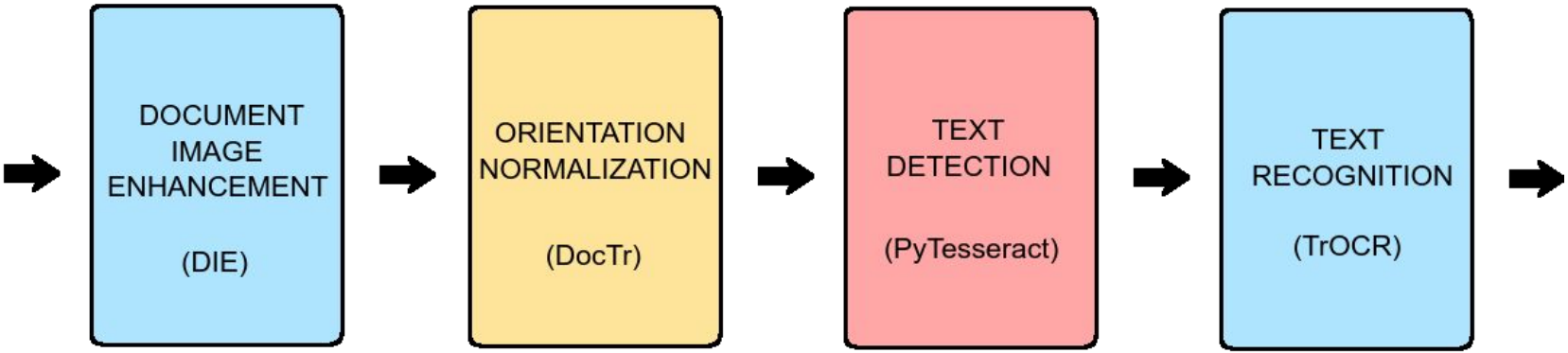
HUN-REN Alfréd Rényi Institute of Mathematics
Artificial Intelligence Group

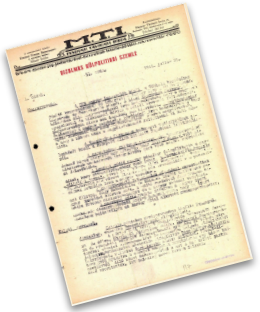
OCR Pipeline

Pipeline: OCR inference

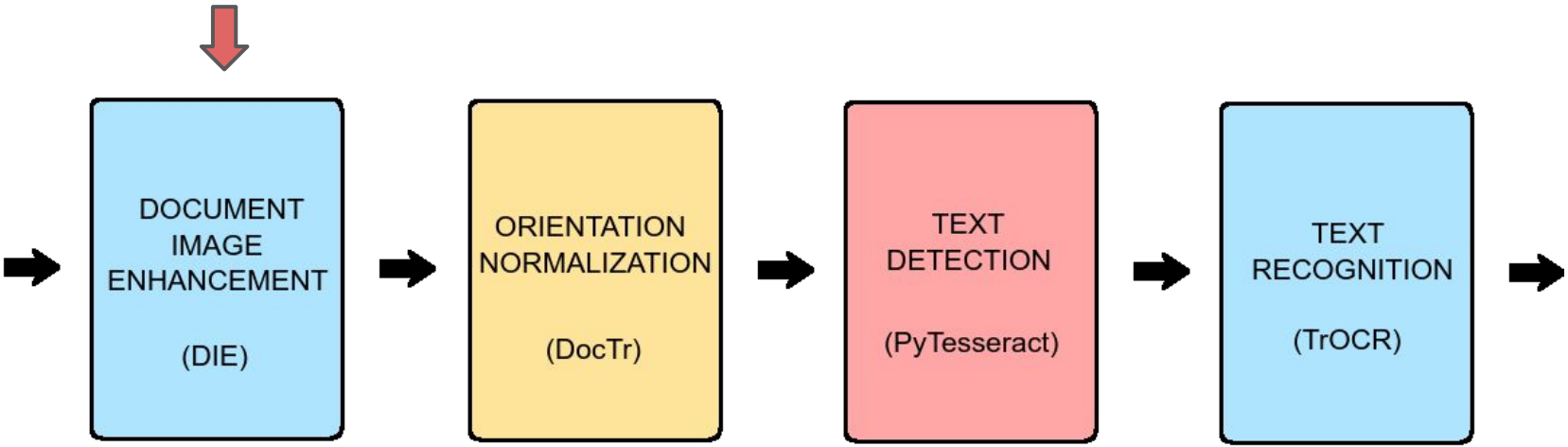


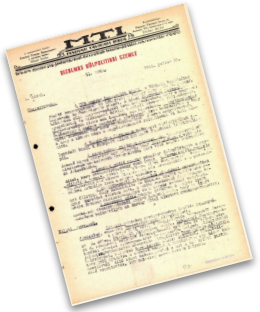
Pipeline: OCR inference



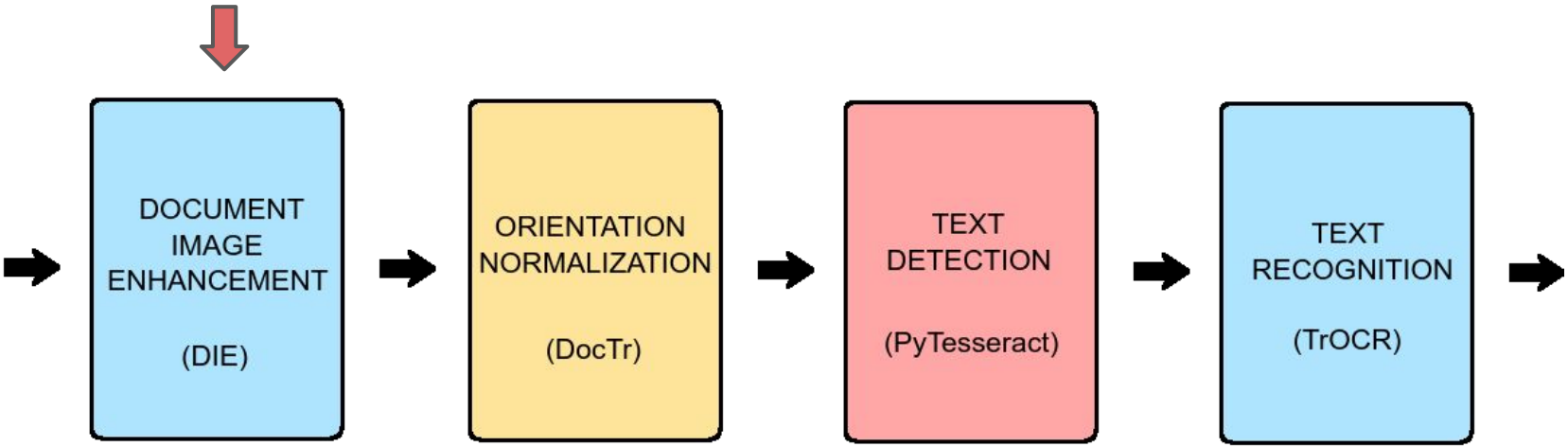


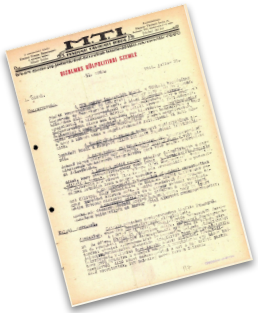
Pipeline: OCR inference



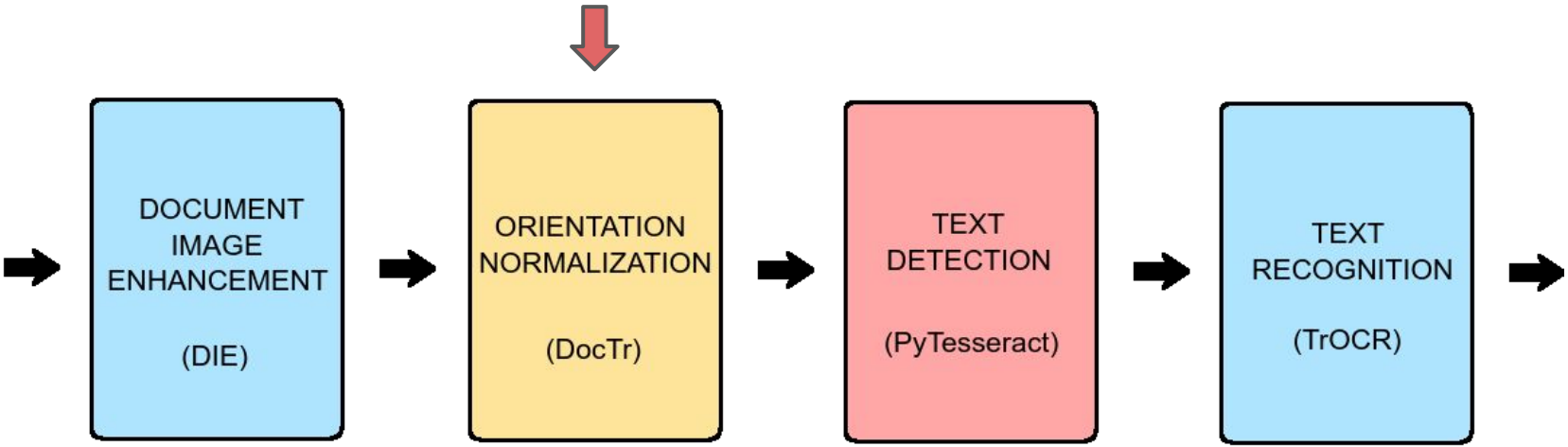


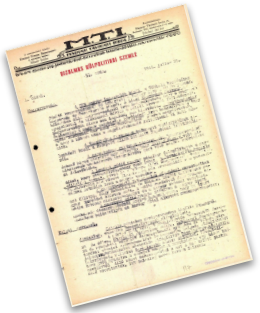
Pipeline: OCR inference



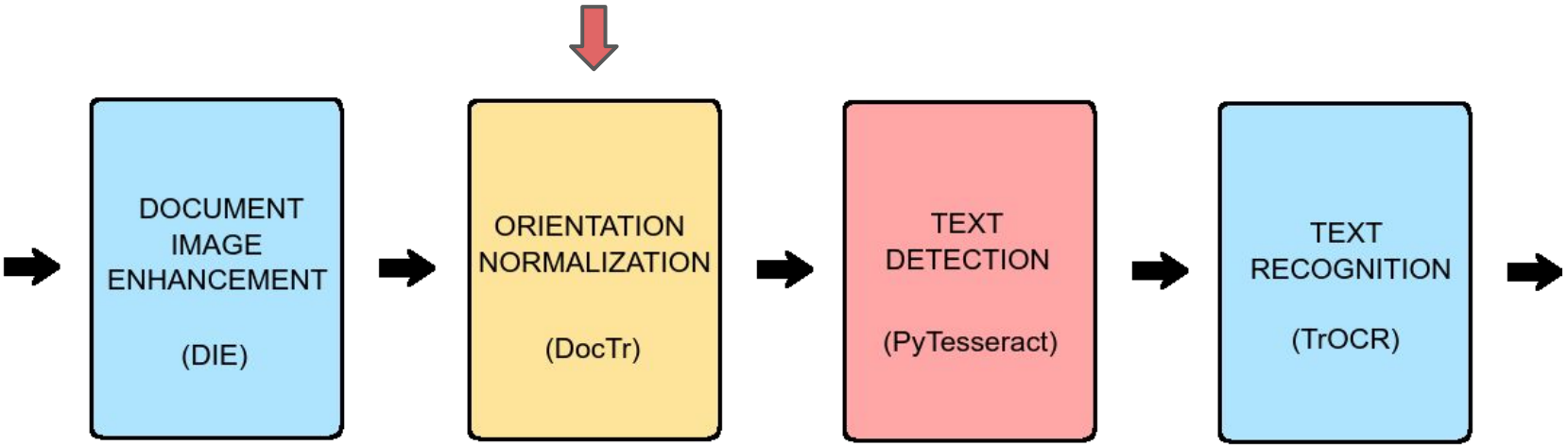


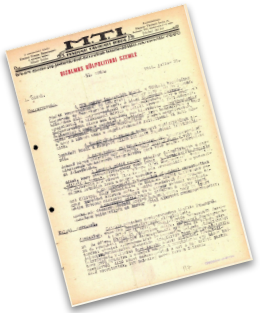
Pipeline: OCR inference

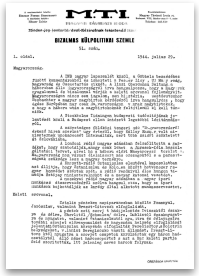




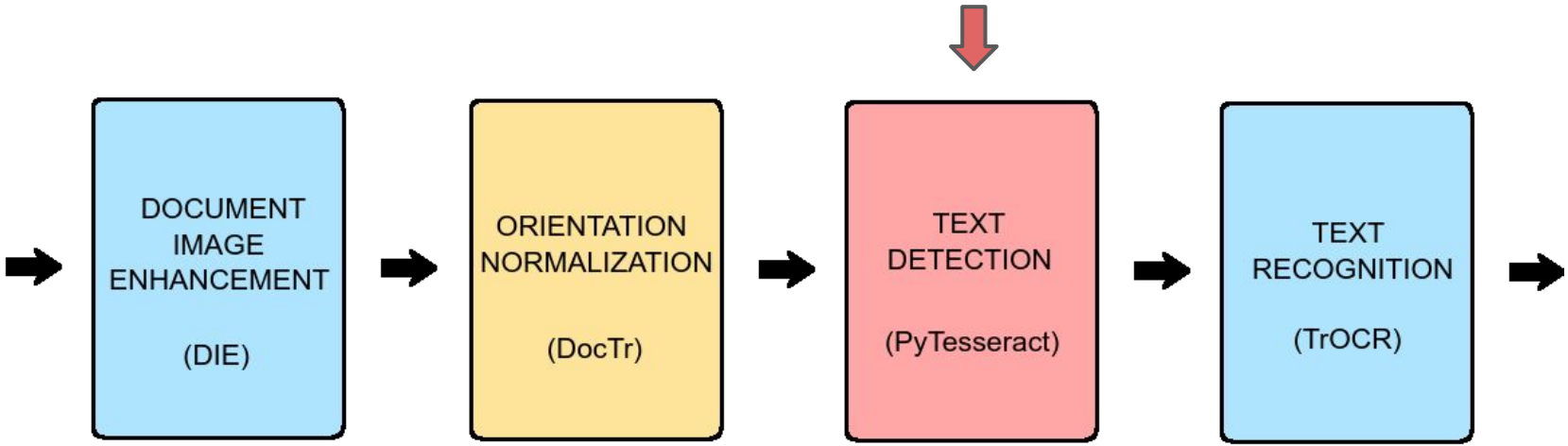
Pipeline: OCR inference

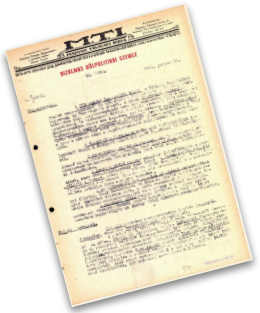


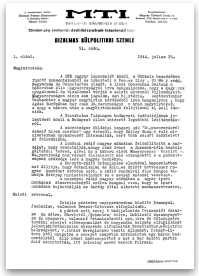




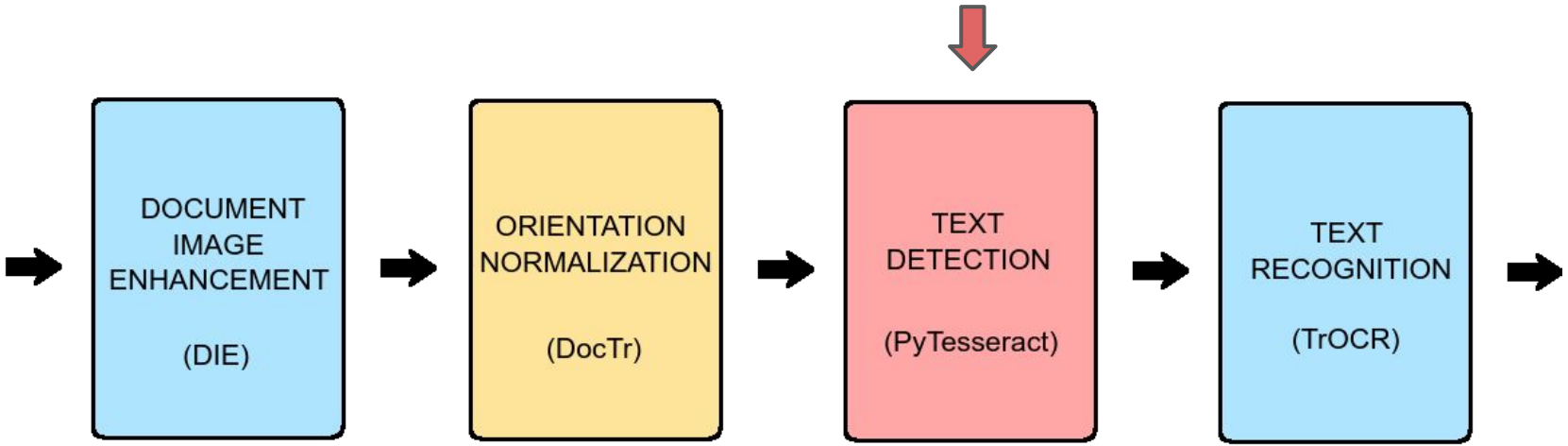
Pipeline: OCR inference

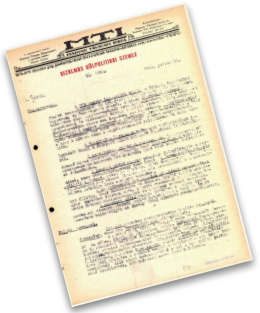


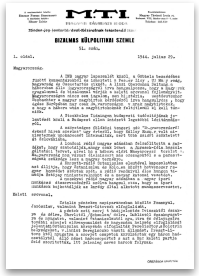


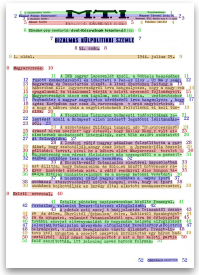


Pipeline: OCR inference





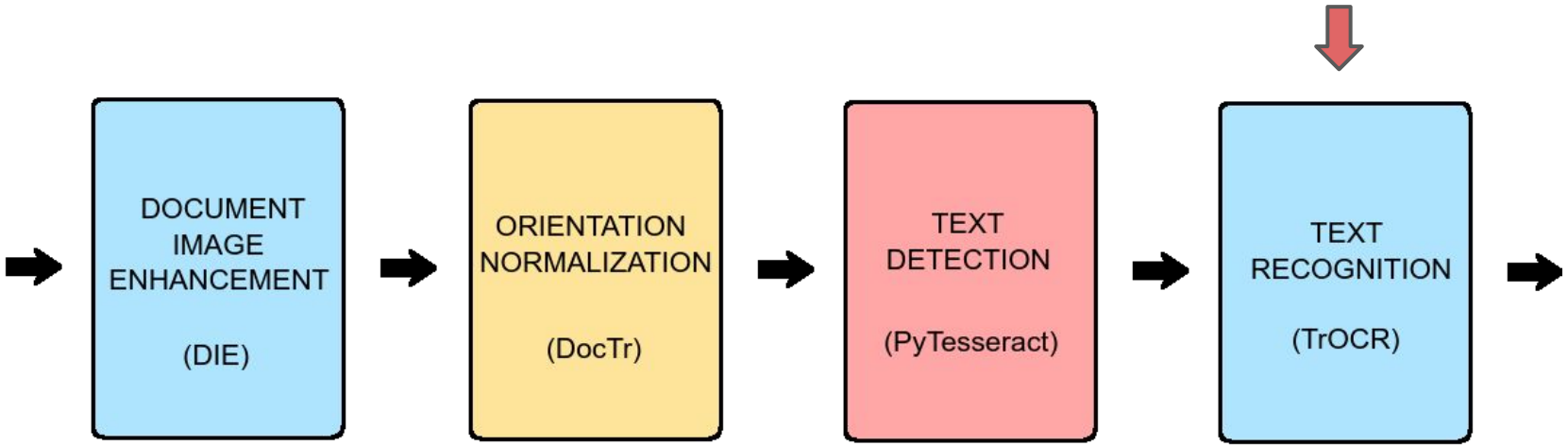


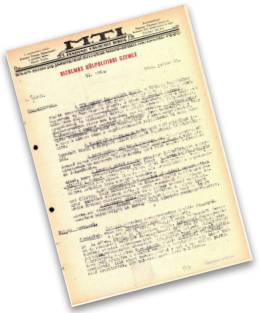


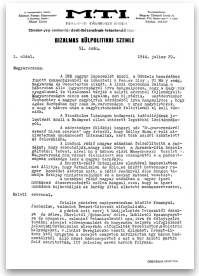
European Commission

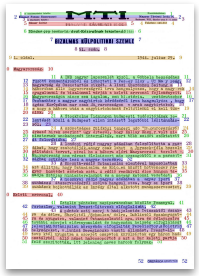
HUN-REN Alfréd Rényi Institute of Mathematics
Artificial Intelligence Group

Pipeline: OCR inference

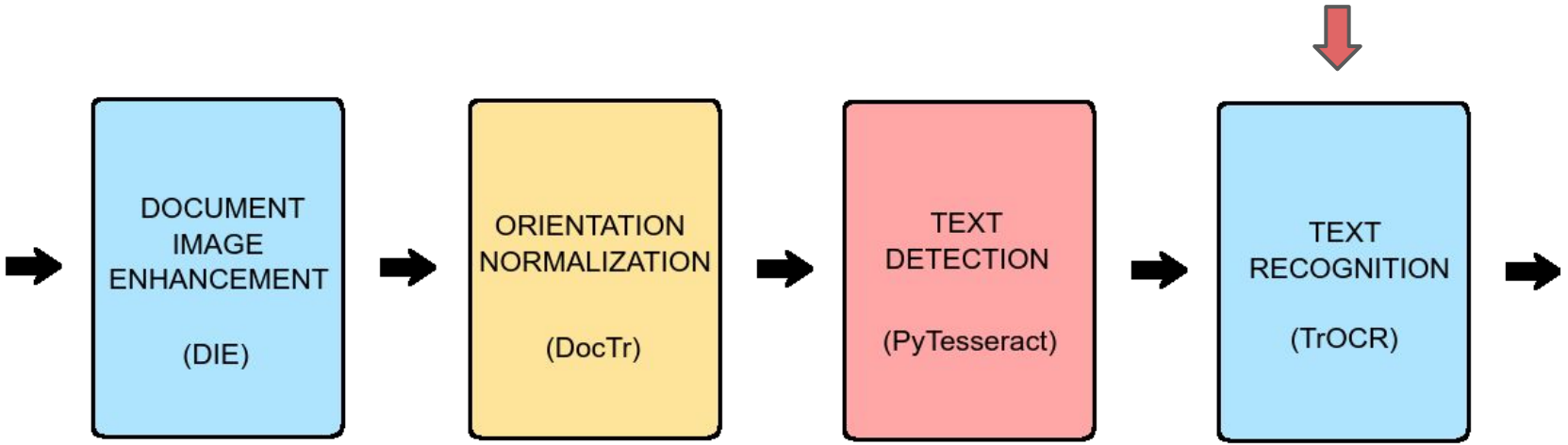


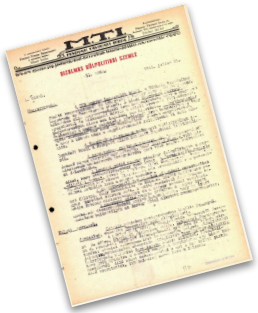


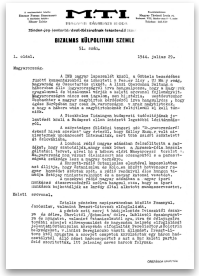


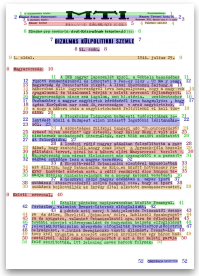


Pipeline: OCR inference





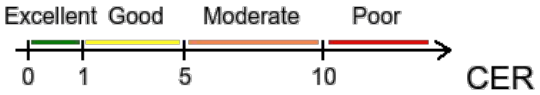




Evaluation

Quantitative result

Performance:



- Character Error Rate (CER)

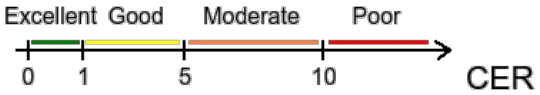
OCR \ Input	Raw document	DIE-cleaned document
Gemini	25.07% [CER]	6.87% [CER]
Google Lens	22.86% [CER]	5.03% [CER]
Our pipeline	**10.59%** [CER] *	**6.46%** [CER]

Inference time:

- Nvidia A-100 server:
  - 8 GPUs (40GB)
  - 2 GPUs:
    - 25 000 document images per day

Quantitative result

Performance:



- Character Error Rate (CER)

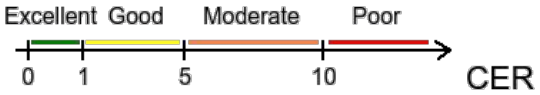
OCR \ Input	Raw document	DIE-cleaned document
Gemini	25.07% [CER]	6.87% [CER]
Google Lens	22.86% [CER]	5.03% [CER]
Our pipeline	**10.59%** [CER] *	**6.46%** [CER]

Inference time:

- Nvidia A-100 server:
  - 8 GPUs (40GB)
  - 2 GPUs:
    - 25 000 document images per day

Quantitative result

Performance:



- Character Error Rate (CER)

OCR \ Input	Raw document	DIE-cleaned document
Gemini	25.07% [CER]	6.87% [CER]
Google Lens	22.86% [CER]	5.03% [CER]
Our pipeline	**10.59%** [CER] *	**6.46%** [CER]

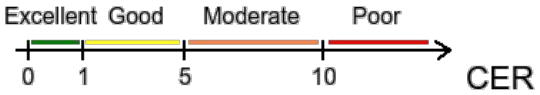
Local model
- **data privacy**

Inference time:

- Nvidia A-100 server:
  - 8 GPUs (40GB)
  - 2 GPUs:
    - 25 000 document images per day

Quantitative result

Performance:



- Character Error Rate (CER)

OCR \ Input	Raw document	DIE-cleaned document
Gemini	25.07% [CER]	6.87% [CER]
Google Lens	22.86% [CER]	5.03% [CER]
Our pipeline	**10.59%** [CER] *	**6.46%** [CER]

Local model
- **data privacy**

Inference time:

- Nvidia A-100 server:
  - 8 GPUs (40GB)
  - 2 GPUs:
    - 25 000 document images per day

Applications

Applications

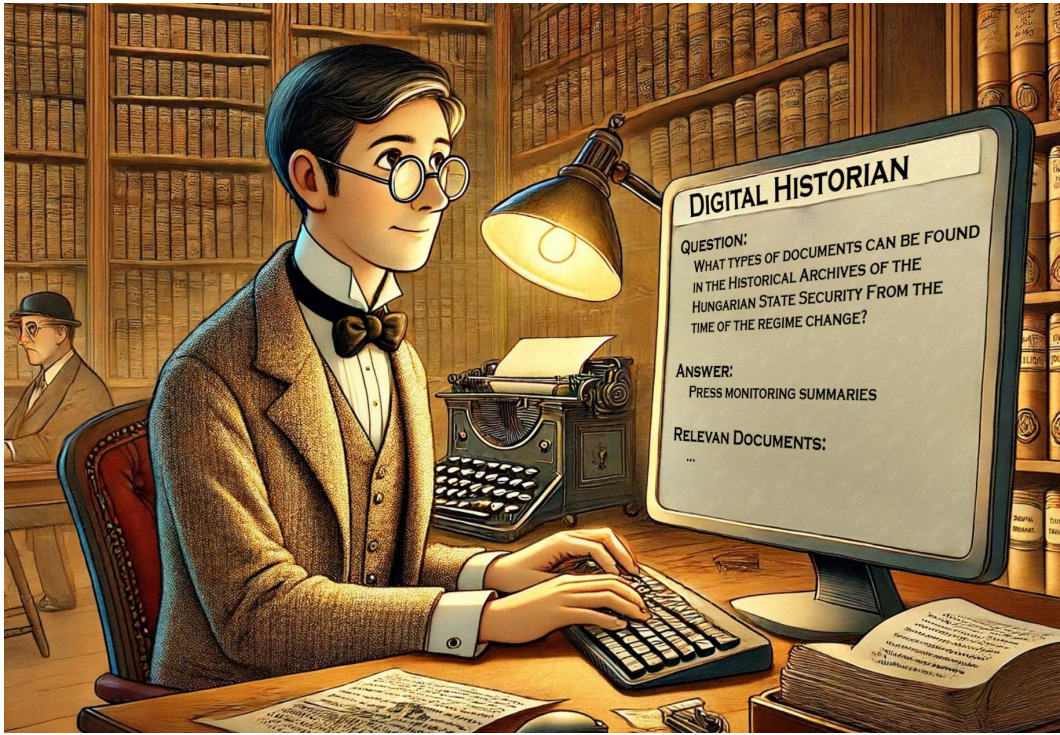
LLMs in Archival Work:

- **Digital Historian**:
  - Retrieval Augmented Generation (RAG)
  - Semantic search in archival database
  - Automatic summarization of documents
  - Dynamic cross-referencing of sources
- **Sensitive Data Removal**:
  - Sensitive and personal information
  - Abstract concepts described in natural language
  - Flexible customization for sensitive data needs
- **Named Entity Recognition (NER)**:
  - Recognition of names, locations, events, etc.
  - Building structured representation (e.g., graphs, tables)
  - Distinguishing between similar or overlapping entities

Applications

LLMs in Archival Work:

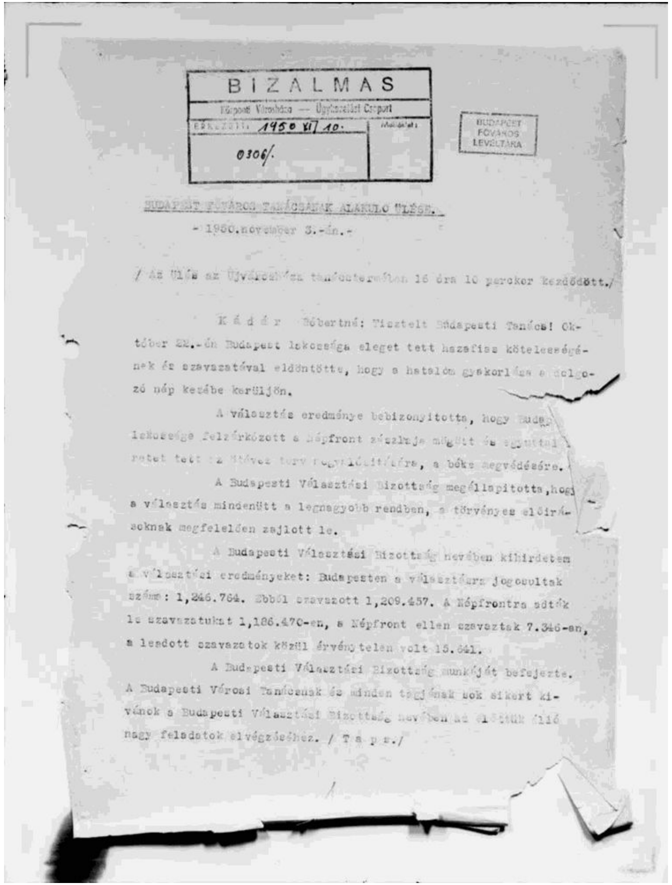
- **Digital Historian**:
  - Retrieval Augmented Generation (RAG)
  - Semantic search in archival database
  - Automatic summarization of documents
  - Dynamic cross-referencing of sources
- **Sensitive Data Removal**:
  - Sensitive and personal information
  - Abstract concepts described in natural language
  - Flexible customization for sensitive data needs
- **Named Entity Recognition (NER)**:
  - Recognition of names, locations, events, etc.
  - Building structured representation (e.g., graphs, tables)
  - Distinguishing between similar or overlapping entities



Applications

LLMs in Archival Work:

- **Digital Historian**:
  - Retrieval Augmented Generation (RAG)
  - Semantic search in archival database
  - Automatic summarization of documents
  - Dynamic cross-referencing of sources
- **Sensitive Data Removal**:
  - Sensitive and personal information
  - Abstract concepts described in natural language
  - Flexible customization for sensitive data needs
- **Named Entity Recognition (NER)**:
  - Recognition of names, locations, events, etc.
  - Building structured representation (e.g., graphs, tables)
  - Distinguishing between similar or overlapping entities



BIZALMAS
Központi Városháza — Ügykezelési Csoport
Érkezett: 1950 XI/ 10.
0306/.

BUDAPEST FŐVÁROS LEVÉLTÁRA

BUDAPEST FŐVÁROS TANÁCSÁNAK ALAKULÓ ÜLÉSE.

- 1950. november 3.-án.-

/ Az ülés az Újvárosháza tanácstermében 16 óra 10 perckor kezdődött./

K á d á r Róbertné: Tisztelt Budapesti Tanács! Október 22.-én Budapest lakossága eleget tett hazafias kötelességének és szavazatával eldöntötte, hogy a hatalom gyakorlása a dolgozó nép kezébe kerüljön.

A választás eredménye bebizonyitotta, hogy Budapest lakossága felzárkózott a Népfront zászlaja mögött és egyúttal [illegible] tett az [illegible] terv megvalósítására, a béke megvédésére.

A Budapesti Választási Bizottság megállapitotta, hogy a választás mindenütt a legnagyobb rendben, a törvényes előirásoknak megfelelően zajlott le.

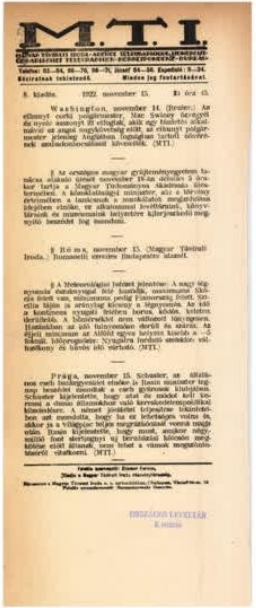
A Budapesti Választási Bizottság nevében kihirdetem a választási eredményeket: Budapesten a választásra jogosultak száma: 1,246.764. Ebből szavazott 1,209.457. A Népfrontra adták le szavazatukat 1,186.470-en, a Népfront ellen szavaztak 7.346-an, a leadott szavazatok közül érvénytelen volt 15.641.

A Budapesti Választási Bizottság munkáját befejezte. A Budapesti Városi Tanácsnak és minden tagjának sok sikert kívánok a Budapesti Választási Bizottság nevében az előttük álló nagy feladatok elvégzéséhez. / T a p s./

Applications

LLMs in Archival Work:

- **Digital Historian**:
  - Retrieval Augmented Generation (RAG)
  - Semantic search in archival database
  - Automatic summarization of documents
  - Dynamic cross-referencing of sources
- **Sensitive Data Removal**:
  - Sensitive and personal information
  - Abstract concepts described in natural language
  - Flexible customization for sensitive data needs
- **Named Entity Recognition (NER)**:
  - Recognition of names, locations, events, etc.
  - Building structured representation (e.g., graphs, tables)
  - Distinguishing between similar or overlapping entities



Raw document image

Washington, november 14. (Reuter.) Az elhunyt corki polgármester, Mac Swiney özvegyét és nyolc asszonyt itt elfogtak, akik egy tüntetés alkalmával az angol nagykövetség előtt, az elhunyt polgármester jelenleg Angliában fogságban tartott növérének szabadonbocsátását követelték. (MTI.)

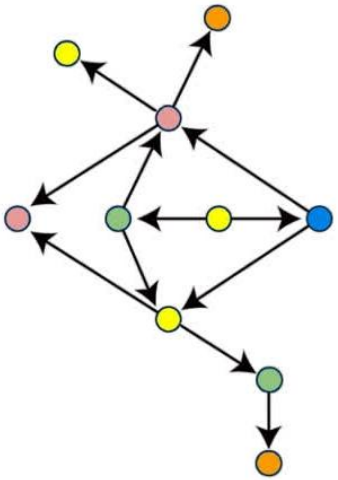
§ Az országos magyar gyűjteményegyetem tanácsa alakuló ülését november 18-án délután 5 órakor tartja a Magyar Tudományos Akadémia üléstermében. A közoktatásügyi miniszter, aki a törvény értelmében a tanácsnak a munkálatok megindulása idejében elnöke, ez alkalommal levéltáraink, könyvtáraink és muzeumaink helyzetére kiterjeszkedő megnyitó beszédet fog mondani.

§ Róma, november 15. (Magyar Távirati Iroda...) Romanelli ezredes Budapestre utazott.

§ A Meteorológiai Intézet jelentése: A nagy légnyomás északnyugat felé huzódik, maximuma Skócia felett van, minimuma pedig Finnország felett. Szicília táján is aránylag kicsi a légnyomás. Az idő a kontinens nyugati felében borus, ködös, keleten derültebb. A hőmérséklet nem változott lényegesen. Hazánkban az idő tulnyomóan derült és száraz. Az éjjeli minimum az Alföld egyes helyein kisebb a —5 foknál. Időprognózis: Nyugatra forduló szelekkel változékony és hűvös idő várható. (MTI.)

Prága, november 15. Schuster, az állásnélküli csehi bankegyesület elnöke is Rasin miniszter tegnap beszédet mondtak a cseh gyárosok klubjában. Schuster kijelentette, hogy utat és módot kell keresni a dunai államokhoz való kereskedelmipolitikai közeledésére. A német jóvátétel teljesítése tekintetében azt mondotta, hogy ha ez lehetséges volna is, akkor is a világpiac teljes megrázkódását vonna maga után. Rasin kijelentette, hogy most amikor négymillió font sterlingnyi uj beruházási kölcsön megkötése előtt állnak, nem lehet a vámok megszüntetéséről vitatkozni. (MTI.)

NER: (persons, locations, dates, ...)



Graph-based data representation

European Commission

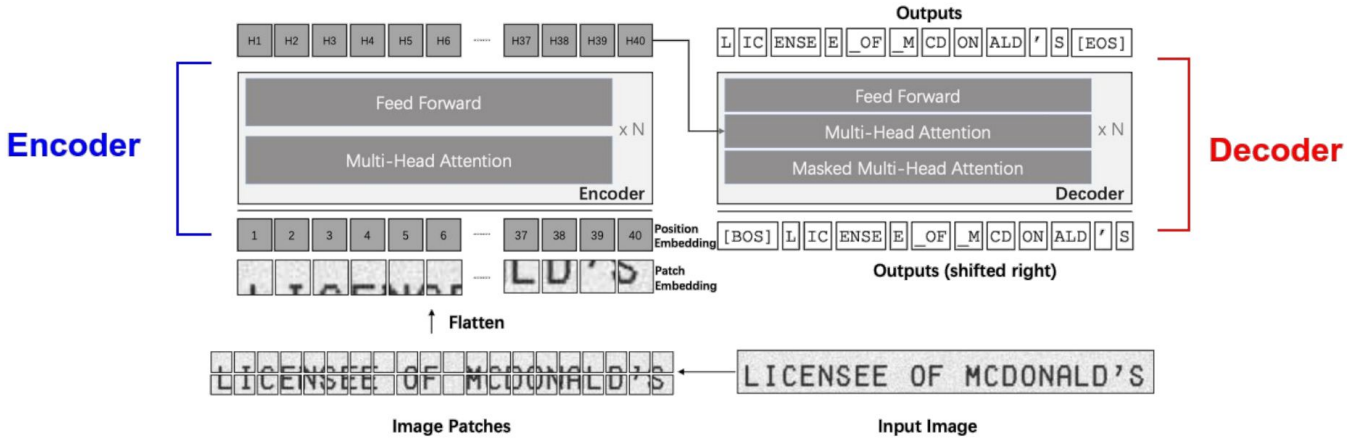
HUN-REN Alfréd Rényi Institute of Mathematics
Artificial Intelligence Group

Thank you for your attention!

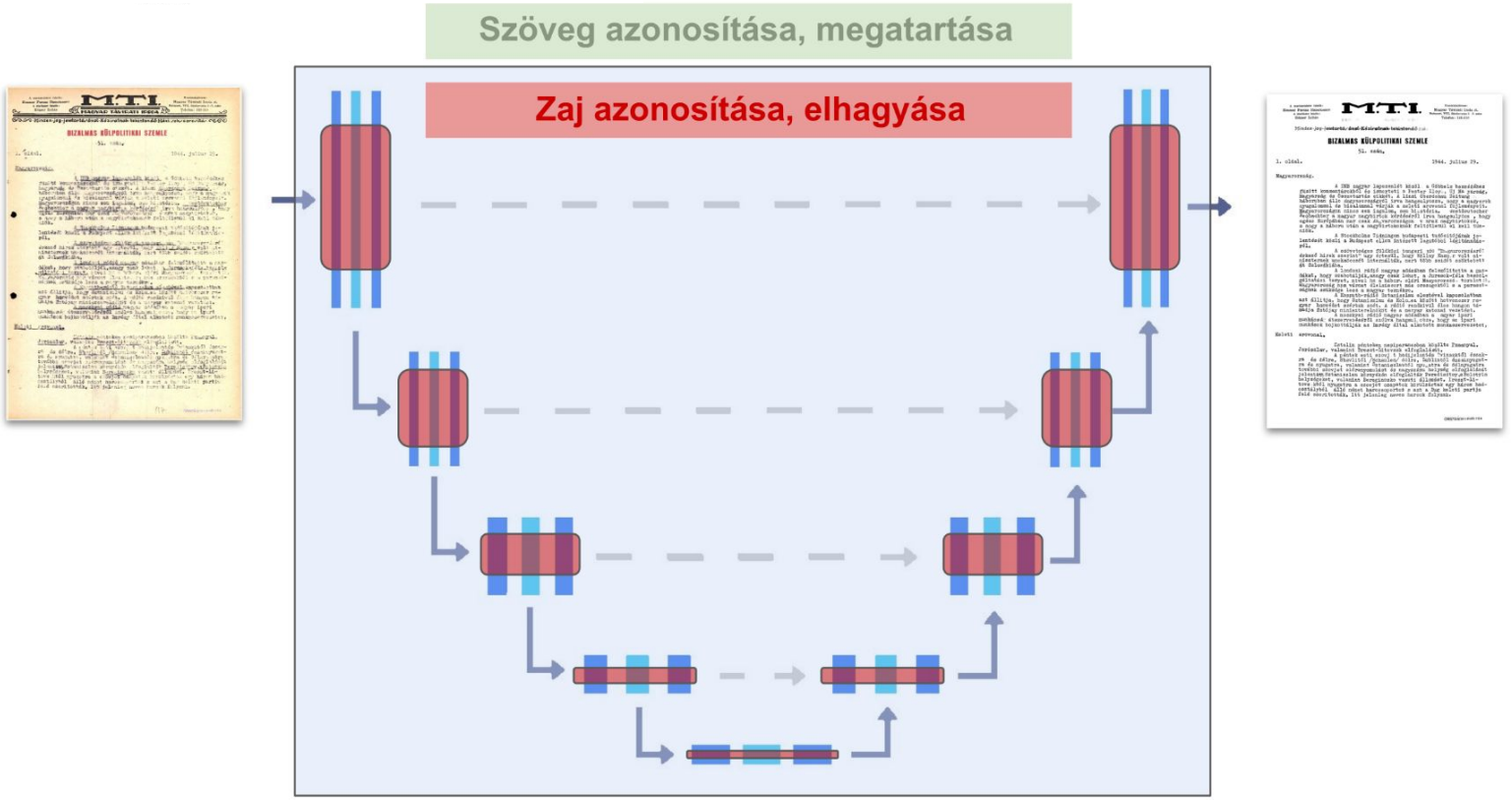
Backup Slides

TrOCR

Text Recognition: TrOCR



U-Net


Szöveg azonosítása, megatartása
Zaj azonosítása, elhagyása

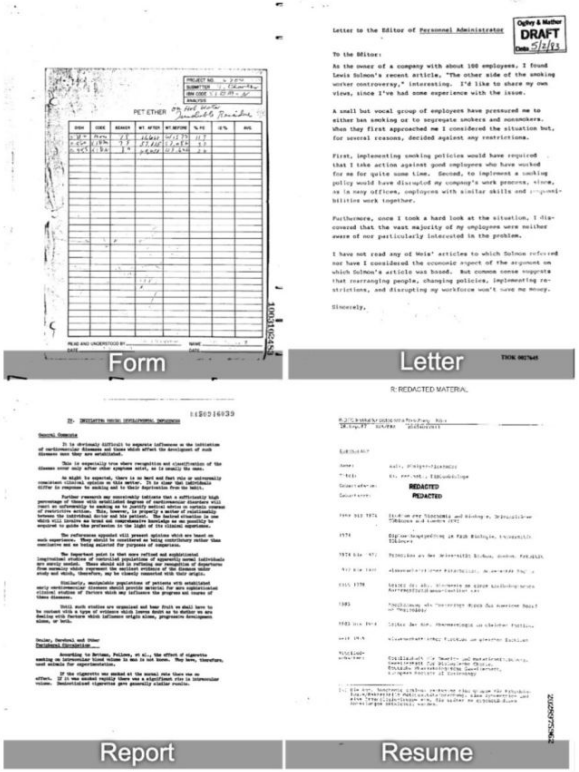
OCR datasets

Scene text

(IIIT5K, SVT, IC03, IC13, SVHN)

Receipts

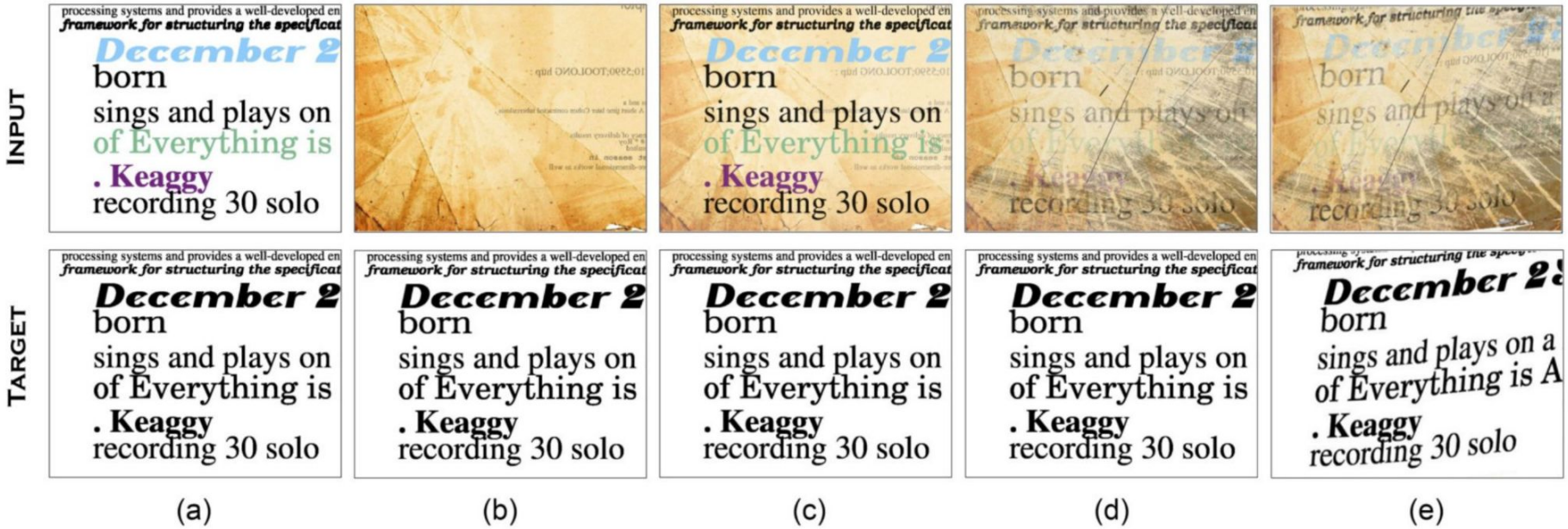
(SROIE)



Documents

(RVL-CDIP)

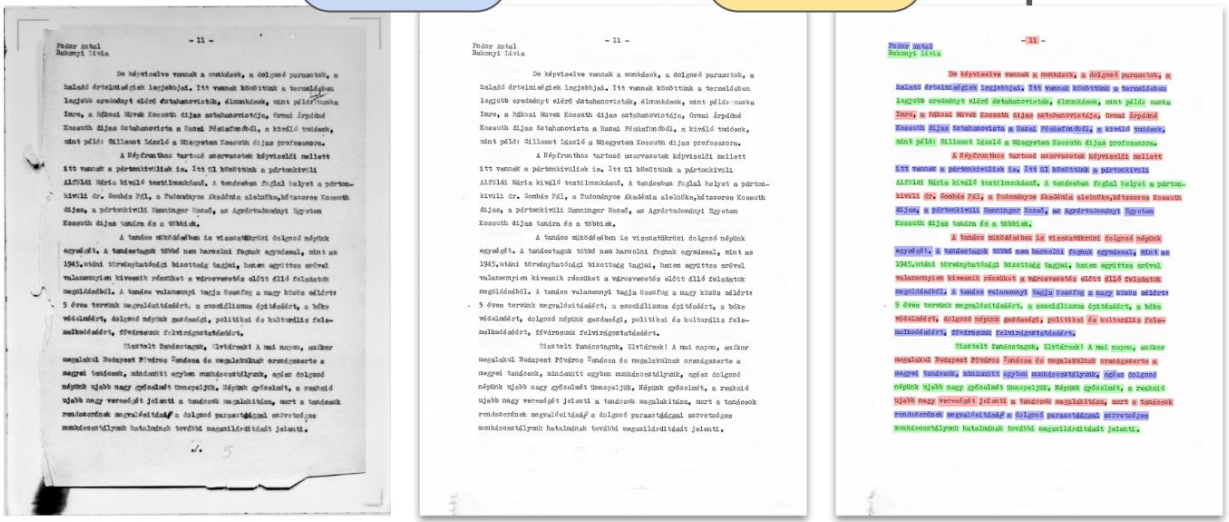
Synthetic Data Generation: Main stages



OCR steps

Document Image Enhancement (DIE)

Text Line Detection (DocTR)



Kossuth díjas Sztahanovista a Hazai Fésüsfonóból, a kiváló tudósok,

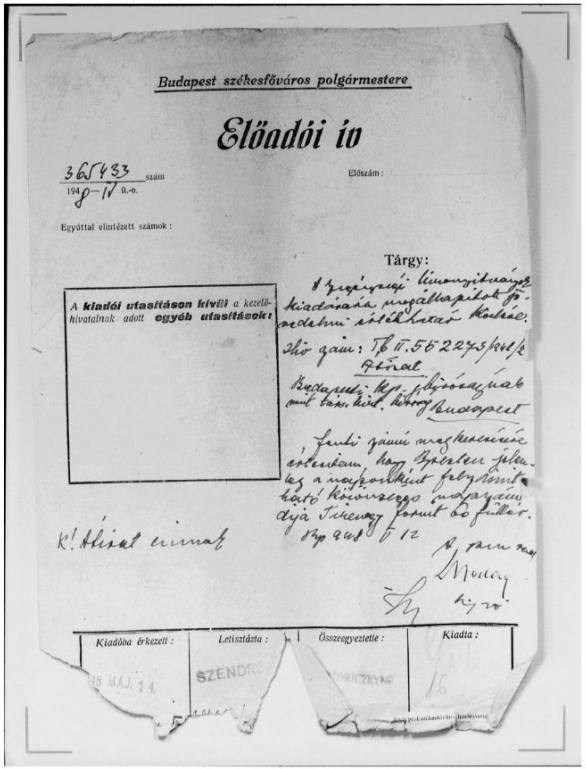
Image

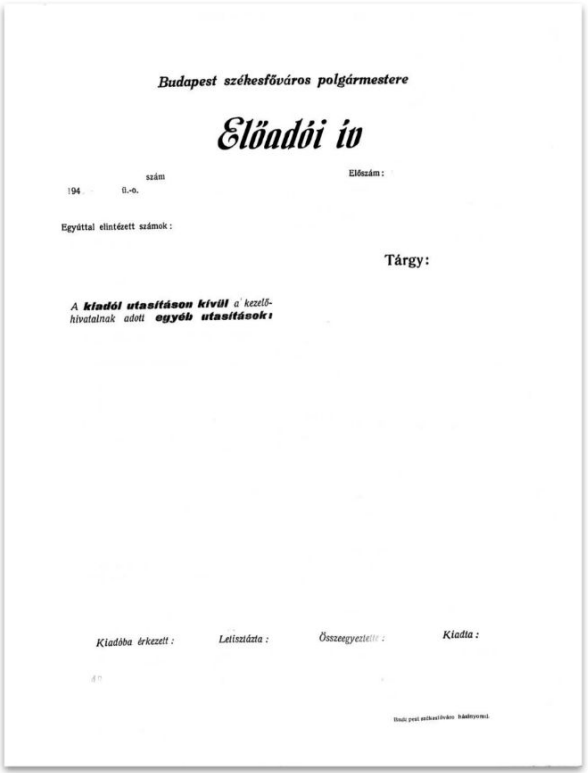
Optical Character Recognition (TrOCR)

Text

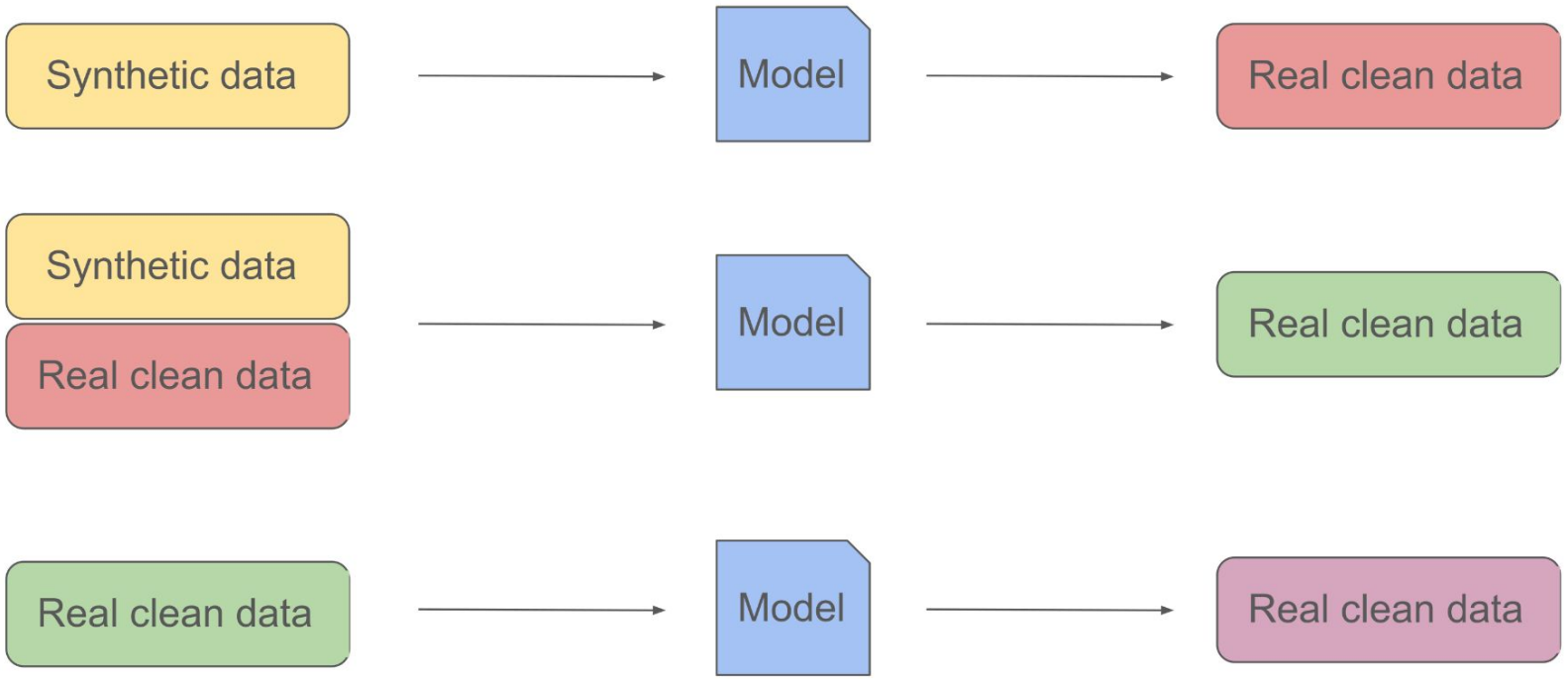
Kossuth díjas Sztahanovista a Hazai Fésüsfonóból, a kiváló tudósok,

DIE: removing handwriting





Bootstrapping: iterative improvement



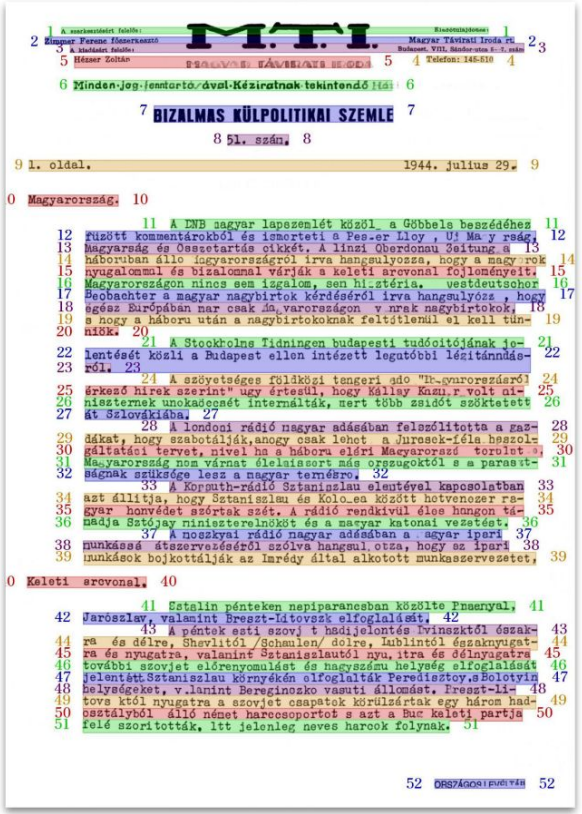
Advantage of DIE

Advantage:

- **Text Detection**:
- **Text Recognition**

Raw document image

Cleaned document image



OCR results on private dataset (ÁBTL)

CER: Character Error Rate
WER: Word Error Rate
NLD: Normalized Levenshtein Distance

Enhance Image Quality

OCR methods	No Preprocess			Simple Preprocess			U-Net (Ours)		
	CER	WER	NLD	CER	WER	NLD	CER	WER	NLD
Tesseract	40.83	67.19	0.407	50.27	97.50	0.504	6.38	28.32	0.061
TrOCR	39.04	62.78	0.389	44.06	87.52	0.443	6.54	30.08	0.063