

Before the LLM Magic Happens:

Clean, Controllable, Reliable Data Extraction with End-to-End
ATR Models

Andy Stauder

Managing Director of read-coop snc



The New York Times.

"All the News That's Fit to Print."

NEW YORK, MONDAY, NOVEMBER 11, 1918. THIRTY-FOUR PAGES.
TWO CENTS

THE WEATHER
Forecast for the City of New York:
Clear; cold; light breeze from west-northwest;
temperature about 60 degrees.

ARMISTICE SIGNED, END OF THE WAR! BERLIN SEIZED BY REVOLUTIONARIES; NEW CHANCELOR BEGS FOR ORDER; GUSTED KAISER FLEES TO HOLLAND

SIN FLEES WITH TREASURY
Hindenburg Absconded
to Be Among Those
In His Party.

ALL ARE HEAVILY ARMED

Armistice Breaks with Rifles as Fighting About All Fronts Ends

ON THEIR WAY TO SEE STEIN

Engineers Not to "Shame,"
Now On Their Way to Berlin

LONDON, Nov. 10.—(By The Associated Press.)—The Kaiser fled from his palace at Spa, Belgium, and was seen by a British aviator flying over the town of Spa, Belgium, according to reports received here today. The Kaiser is believed to have fled to Holland, where he will be met by the Dutch Government, it was learned today. The Kaiser is believed to have fled to Holland, where he will be met by the Dutch Government, it was learned today. The Kaiser is believed to have fled to Holland, where he will be met by the Dutch Government, it was learned today.

BREITEN TROOPS JOIN RESIST.
Reds Still Building in Which Officers Vainly Resist.

THRONES DEMAND REPUBLIC

Revolutionary Flag on Reich Palace—Crown Prince's Palace Seized

GERMANY SET IN FLAME

General Staff in Berlin

LONDON, Nov. 10.—(By The Associated Press.)—The Kaiser fled from his palace at Spa, Belgium, and was seen by a British aviator flying over the town of Spa, Belgium, according to reports received here today. The Kaiser is believed to have fled to Holland, where he will be met by the Dutch Government, it was learned today. The Kaiser is believed to have fled to Holland, where he will be met by the Dutch Government, it was learned today.

British Chancellor Appeals to All Germans
To Help Him Save Fatherland From Anarchy

WAR LOSERS AT 6 O'CLOCK THIS MORNING

The State Department in Washington Made the Announcement at 2:45 o'clock.

ARMISTICE WAS SIGNED IN FRANCE AT MIDNIGHT

Terms include Withdrawal from Alsace-Lorraine, Disarming and Demobilization of Army and Navy, and Occupation of Strategic Naval and Military Ports

WASHINGTON, Monday, Nov. 11, 7:48 A. M.—The armistice between Germany, on the one hand, and the allied Governments of Great Britain, the United States, on the other, has been signed. The State Department announced at 2:45 o'clock this morning that Germany had signed.

The department's announcement simply said: "The armistice has been signed." The world war will end this morning at 6 o'clock, Washington time, 11 o'clock Paris time. The armistice was signed by the German representatives at midnight.

This announcement was made by the State Department at 2:45 o'clock this morning. The announcement was made verbally by an official of the State Department in this form: "The armistice has been signed. It was signed at 6 o'clock A. M. Paris time, [midnight, New York time] and hostilities will cease at 11 o'clock this morning, Paris time, (6 o'clock, New York time)." The terms of the armistice, it was assumed, would not be made public until later. Military men here, however, regard it as certain that they include:

Immediate retirement of the German military forces from France, Belgium, and Alsace-Lorraine.

Disarming and demobilization of the German armies.

Occupation by the allied and American forces of such strategic points in Germany as will make impossible a renewal of hostilities.

Delivery of part of the German High Seas Fleet and a certain number of submarines to the allied and American navies.

Disarmament of all other German naval vessels.

Begin War on Your Terms, O
Paris, if You Desire.

I found many new supports in the field for my writing on. You seem like someone the New has touched, the awakening of the people, which is the first awareness of the people's existence. The political organizations should not forget the people. The first step is the first step of an awakening, the first step of a people, and they should be awakened. They are all the products of the first step and they are the first step of the first step.

GENERAL STRIKE IS BEGUN
Burgemeister and Police Submit—War Office Now Under Socialist Control.

[illegible]

These features, which are standard fully equipped, include up the Newer than that. It is not to be left off of general construction, especially with such great. However, the only one left, the subject of the lighting is good, even as compared to the other, which is the best.

ON THEIR WAY TO DIE STEES

Belgians Tell to Them, "Are You On Your Way to Paris?"

The Solution

Clean, controllable, reliable data extraction* with end-to-end ATR models.
*and information extraction

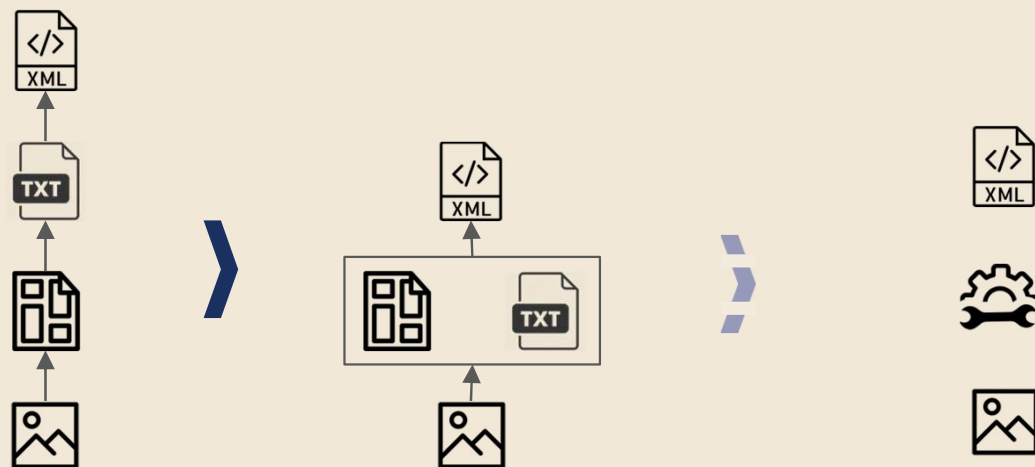
What are data and information?



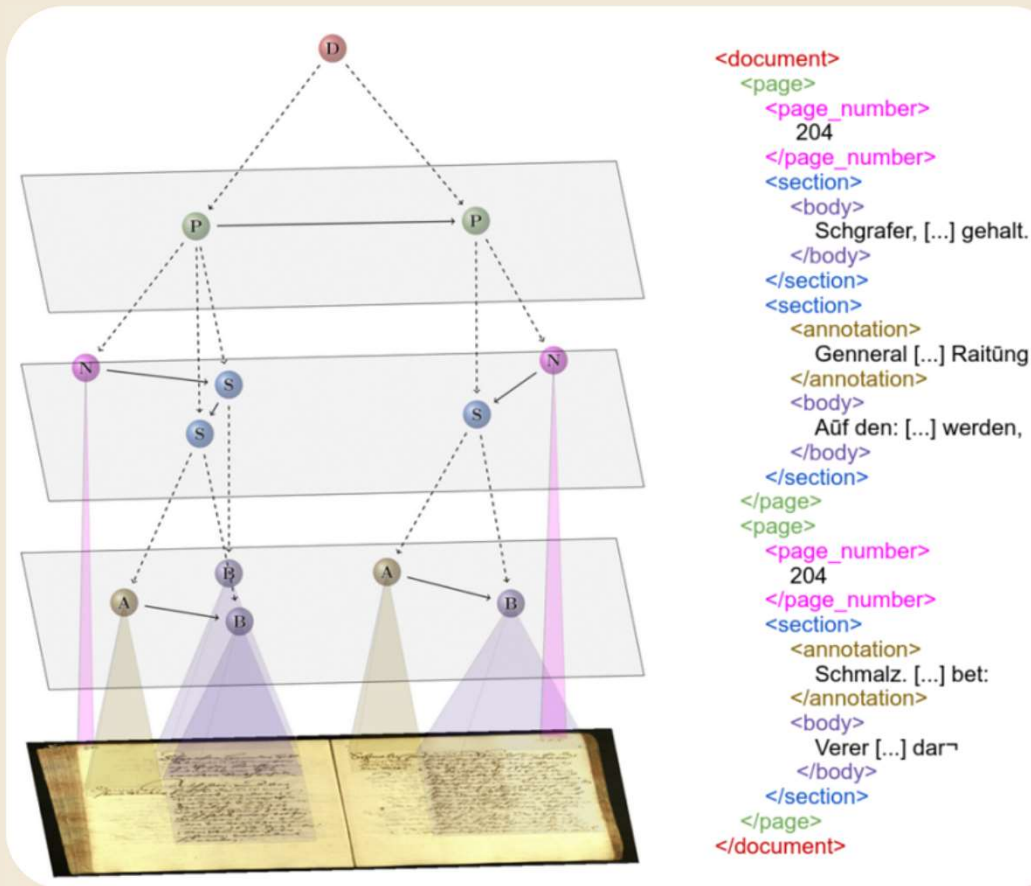
© cngcoins.com

Data: Measured (series of) values
Information: Distinguishable data
Knowledge: Linked information

What are End-to-End ATR Models?



How do they work? DAN Example



Architecture

FCN encoder

visual features

Transformer decoder

characters + layout

Mutual attention/feature sharing

Process

Transfer learning with curriculum

Simple synthetic pages

Gradually transition to real data

Why use End-to-End Models?

No propagation errors

Complex reading orders (rotated, inter-line, tabular)

Nested elements (layout and named entities)

Easier maintenance

Faster

More accurate

Trainable on mid-range hardware

More well-behaved than LLMs

Example: Basque Student Essays

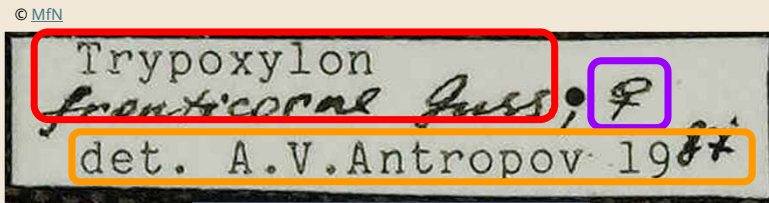
[illegible]

(Symbolic image)

CER: 7% \rightarrow 3%

Layout accuracy: 45% → 90%

Example: Specimen Labels



- Region 1
- 1 Trypoxylon
 - 2 fronticorne Guss., ♀
 - 3 det. A. V. Antropov 1987



Specimen	Sex	Determination
Trypoxylon fronticorne Guss.	♀	det. A.V. Antropov 1987

CER: 27% → 8% (mixed)

NER: F1: 0.96 → 0.94

Labelling Newspapers: Transcription and Labelling



Transcribe text in reading order

Assign bottom level region tags

Assign NER Tags

- 1 Average Daily Circulation For
- 2 Week Ending
- 3 Nov. 15th
- 4 11,448

Region 4 - heading-established

1 ESTABLISHED 1870

Region 5 - heading-date

1 NEW BRITAIN, CONNECTICUT, TUESDAY, NOVEMBER 18, 1924.-EIGHTEEN PAGES.

Region 6 - heading-price

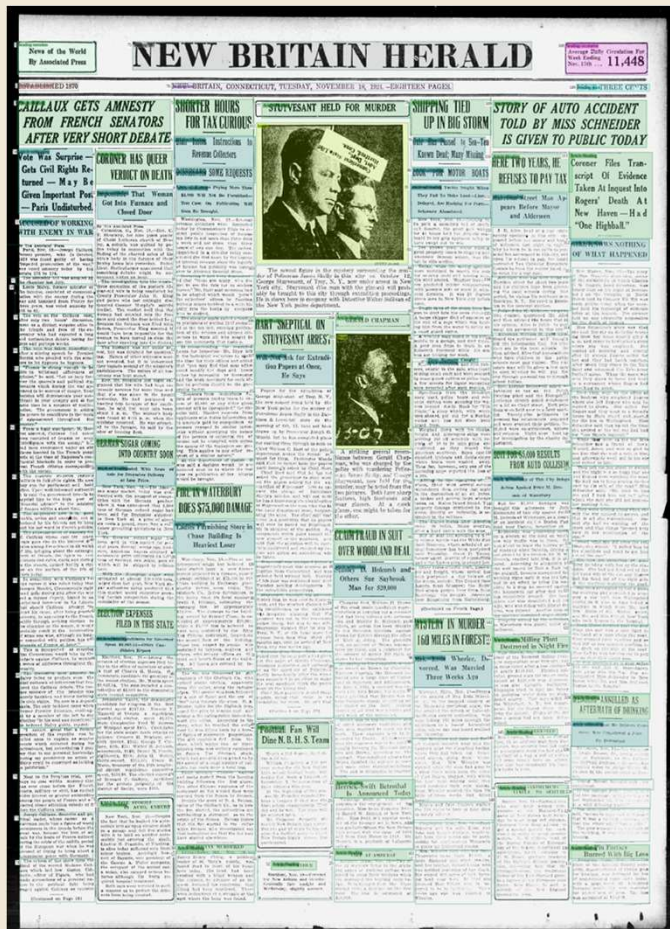
1 PRICE THREE CENTS

Region 7 - Article-Heading

- 1 CAILLAUX GETS AMNESTY
- 2 FROM FRENCH SENATORS

Structured Output

Input



Output

```
1 <Page>
2
3 <heading>
4   <heading-top>
5     <heading-metadata>
6       News of the World
7       By Associated Press
8     </heading-metadata>
9     <heading-title>
10      NEW BRITAIN HERALD
11    </heading-title>
12    <heading-circulation>
13      Average Daily Circulation For
14      Week Ending
15      <tag_date>Nov. 15th...</tag_date>
16      11,448
17    </heading-circulation>
18  </heading-top>
19  <heading-bottom>
20    <heading-established>
21      ESTABLISHED <tag_date>1876</tag_date>
22    </heading-established>
23    <heading-date>
24      NEW BRITAIN, CONNECTICUT, TUESDAY, <tag_date>NOVEMBER 18, 1924</tag_date>...EIGHTEEN PAGES.
25    </heading-date>
26    <heading-price>
27      PRICE THREE CENTS
28    </heading-price>
29  </heading-bottom>
30 </>
```

Next Steps

General synthetic page generator

Universal data format

Tackle sequence length constraints

Write back positional data

Training and inference in Transkribus for all



Unlock every doc.