

# Chunk-based approach

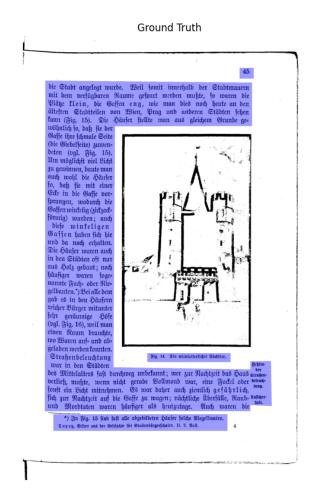
NLP - Praktikum

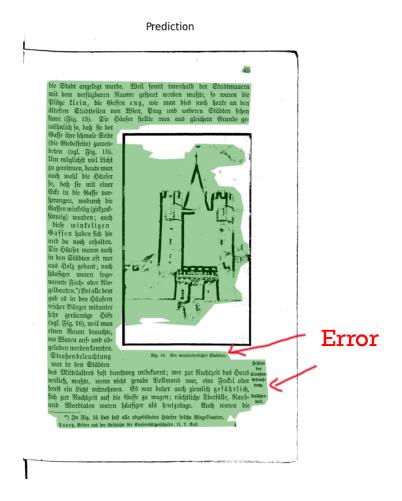
Alzbeta Hrabosova



# Manual data labeling

Error score per page
 error words / total words







#### **Ground Truth**

#### Prediction

194

ganz Europa, auf welcher die durch Napoleon in Verwirrung gebrachten staatlichen Verhältnisse neu geordnet wurden. Die Oberleitung hatte bei diesen Verhandlungen der einflußreichste Staatsmann, den Europa damals besaß, der österreichische Minister Metternich, der kurz zuvor zur Belohnung für seine Verdienste um die Machtstellung Österreichs in den Fürstenstand erhoben worden war. Österreich erhielt durch den Wiener Kongreß im wesentlichen den Umfang, den es heute noch besitzt; von den Erwerbungen, die es zur Zeit Napoleons gemacht hatte, blieben ihm insbesondere Salzburg und Venezien, wodurch die Monarchie, da auch die Lombardei österreichisch war, in zweckmäßiger Weise abgerundet (arrondiert) wurde. Im deutschen Bunde, der an die Stelle des deutschen Reiches trat, führte Österreich den Vorsitz.

Zu Ehren der in Wien versammelten Monarchen und Würdenträger fanden rauschende, glanzvolle Festlichkeiten statt, die lange im Gedächtnisse der Zeitgenossen fortlebten. Viele derselben wurden in Laxenburg gefeiert, dem Lieblingssitze des Kaisers Franz, der sich daselbst in getreuer Nachbildung mittelalterlicher Burgen die "Franzensburg" hatte erbauen lassen (s. die Abbildung Fig. 59). Eines dieser Feste verherrlichte auch der größte Meister der Töne, den die Welt damals besaß, Ludwig van Beethoven, durch die Aufführung eines seiner Tonwerke, der "Schlacht bei Vittoria", (Bei Vittoria in Spanien hatte der englische Feldherr Wellington, welcher ebenfalls dem Kongresse beiwohnte, 1812 einen glänzenden Sieg über die Franzosen erfochten.)

Unter der väterlichen Regierung des "guten Kaisers Franz" erholte sich Österreich rasch von den Wunden, welche ihm die Napoleonischen Kriege geschlagen. Ackerbau und Viehzucht, Handel und Gewerbe, Wissenschaft und Kunst gelangten zu neuer Blüte.

Für die Landwirtschaft war besonders die Gründung von landwirtschaftlichen Schulen, dann von Ackerbaugesellschaften und landwirtschaftlichen Vereinen, für das Gewerbe die Errichtung der technischen
Institute in Wien, Prag, Graz u. s. w. und die Veranstaltung von Gewerbeausstellungen wichtig. Erzherzog Johann, der Bruder des Kaisers,
selbst Eigentümer mehrerer Eisenwerke in Steiermark, bewog die Besitzer
der Eisenwerke, welche die unerschöpflichen Spateisensteinlager des steirischen Erzberges ausbeuteten, die noch bestehende Vordernberger
Eisenwerksgesellschaft zu gründen, so daß nun die Gewinnung und Versendung des Eisens im großen betrieben werden konnte. Der Koblenberghau, im Zusammenhang damit das Maschinen- und Eisenbahnwesen
begannen sich in dieser Zeit zu entwickeln. Die erste Eisenbahn
in Österreich war die von Budweis nach Linz (1825), welche aber anfangs
nur eine Pferdebahn war. Auch die großartigen Kunststraßen über den
Loibl-Paß und das Stifser Joch entstanden in dieser Zeit. Kaiser

134

ganz Europa, auf welcher die durch Napoleon in Verwirrung gebrachten staatlichen Verhältnisse neu geordnet wurden. Die Oberleitung hatte bei diesen Verhandlungen der einflußreichste Staatsmann, den Europa damals besaß, der österreichische Minister Metternich, der kurz zuvor zur Belohnung für seine Verdienste um die Machtstellung Österreichs in den Fürstenstand erhoben worden war. Österreich erhielt durch den Wiener Kongreß im wesentlichen den Umfang, den es heute noch besitzt; von den Erwerbungen, die es zur Zeit Napoleons gemacht hatte, blieben ihm insbesondere Salzburg und Venezien, wodurch die Monarchie, da auch die Lombardei österreichisch war, in zweckmäßiger Weise abgerundet (arrondiert) wurde. Im deutschen Bunde, der an die Stelle des deutschen Reiches trat, führte Österreich den Vorsitz.

Zu Ehren der in Wien versammelten Monarchen und Würdenträger fanden rauschende, glanzvolle Festlichkeiten statt, die lange im Gedächtnisse der Zeitgenossen fortlebten. Viele derselben wurden in Laxenburg gefeiert, dem Lieblingssitze des Kaisers Franz, der sich daselbst in getreuer Nachbildung mittellaterlicher Burgen die "Franzensburg" hatte erbauen lassen (s. die Abbildung Fig. 59). Eines dieser Feste verherrlichte auch der größte Meister der Töne, den die Welt damals besaß, Ludwig van Beethoven, durch die Aufführung eines seiner Tonwerke, der "Schlacht bei Vittoria". (Bei Vittoria in Spanien hatte der englische Feldherr Wellington, welcher ebenfalls dem Kongresse beiwohnte, 1812 einen glänzenden Sieg über die Franzosen erfochten.)

Unter der väterlichen Regierung des "guten Kaisers Franz" erholte sich Österreich rasch von den Wunden, welche ihm die Napoleonischen Kriege geschlagen. Ackerbau und Viehzucht, Handel und Gewerbe, Wissenschaft und Kunst gelangten zu neuer Blüte.

Für die Landwirtschaft war besonders die Gründung von landwirtschaftlichen Schulen, dann von Ackerbaugesellschaften und landwirtschaftlichen Vereinen, für das Gewerbe die Errichtung der technischen
Institute in Wien, Prag, Graz u. s. w. und die Veranstallung von Gewerbeausstellungen wichtig. Erzberzog Johann, der Bruder des Kaisers,
selbst Eigentümer mehrerer Eisenwerke in Steiermark, bewog die Besitzer
der Eisenwerke, welche die unerschöpflichen Spateisensteinlager des steirischen Erzberges ausbeuteten, die noch bestehende Vordernberger
Eisenwerksgesellschaft zu gründen, so daß nun die Gewinnung und Versendung des Eisens im großen betrieben werden konnte. Der Koblenbergbau, im Zusammenhang damit das Maschinen- und Eisenbahnwesen
begannen sich in dieser Zeit zu entwickeln. Die erste Eisenbahn
in Österreich war die von Budweis nach Linz (1825), welche aber anfangs
nur eine Pferdebahn war. Auch die großartigen Kunststraßen über den
Loibl-Paß und das Stiffser Joch entstanden in dieser Zeit. Kaiser

No Errors

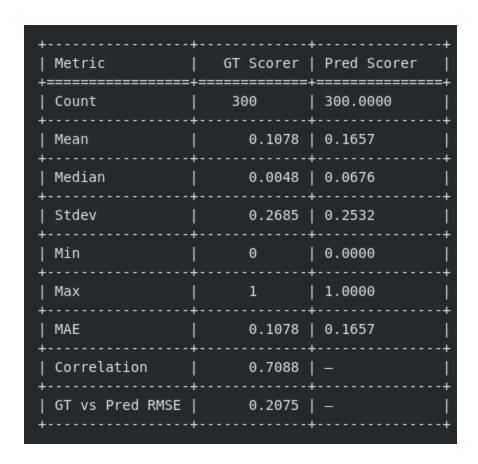


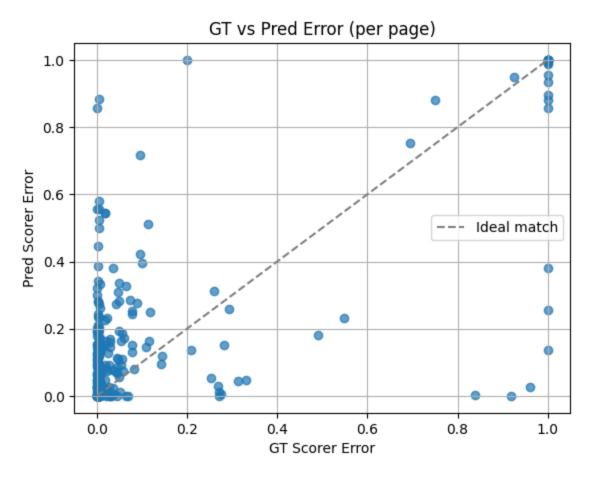
## Chunks-size/Overlap-size experiments

Chunk size	Overlap size	Mean		Std. de	v.	Correlation
		GT	PR	GT	PR	
10	5	0.227	0.503	0.371	0.308	0.81
10	7	0.227	0.514	0.371	0.309	0.799
15	7	0.227	0.328	0.371	0.386	0.851
15	10	0.227	0.328	0.371	0.402	0.833
20	15	0.227	0.278	0.371	0.398	0.879
20	15	0.227	0.284	0.371	0.400	0.877

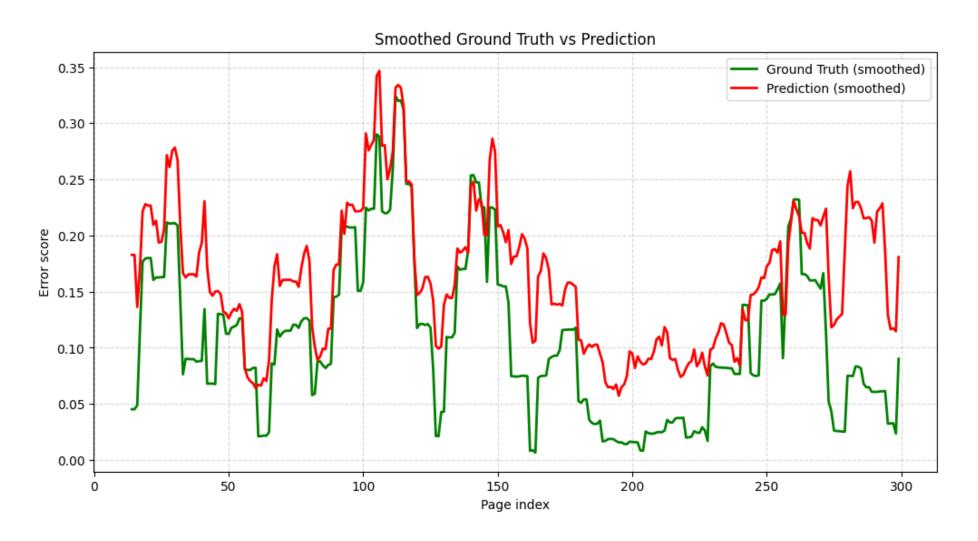
30.06.2025 Alzbeta Hrabosova











## **Supervised OCR Error Detection**

Presented by: Arslan Ahmed

## **Approach**



### **Word-level labeling**

Label each word in input line as "error", "correct", or "unmatched"



## **Tokenization and Model Training**

Train a token classification model to predict error/correct labels



#### **Evaluation**

Assess model performance on ManualGTScorer

## How to label each word?

#### Overview:

- For each input file, we compare each line to its corresponding ground truth (GT) line.
- o If we find an exact line match, all words in that line are labeled as "correct".
- For lines without exact matches, we proceed with contextual (partial) matching to assign word-level labels.

#### **Labeling Rules:**

- 1. Exact Marginalia or Paragraph Match:
- All words in an input line that exactly matches a marginalia or paragraph line in GT are labeled correct.

## How to label each word?

#### 2. Contextual Alignment (Partial Match):

- For lines without an exact match, we perform word-by-word comparison using context from the ground truth paragraph.
- Context: An input word is labeled 'correct' if it has an exact match in the ground truth line,
   along with a match in either its previous or next word.

#### 3. Marginalia Phrase Inside Line:

 Words in input lines that match a marginalia phrase (from GT) within a paragraph line are labeled error.

#### 4. Unmatched Words:

Words not aligned to any GT or marginalia are labeled unmatched.

```
1 0001: We(unmatched) | te(unmatched)
1 0002: O(unmatched)
1 0003: S(unmatched)
1 0004: 1(unmatched)
1 0005: 16(correct)
1 0006: Rückzug(error) | der(error) | wichen(correct) | zurück(correct) | und(correct) | ließen(correct) | sich(correct) | in(correct) | keine(correct)
| offene(correct) | Schlacht(correct) | ein(correct) | sondern(correct)
1 0007: Skythen(correct)
1 0008: verwüsteten(correct) | nur(correct) | das(correct) | Land(correct) | und(correct) | verschütteten(correct) | die(correct) | Brunnen(correct) |
bis(correct) | den(correct)
1 0009: Persern(correct) | die(correct) | Lebensmittel(correct) | ausgingen(correct) | und(correct) | Darius(correct) | den(correct) | Befehl(correct)
| zum(correct)
1 0010: Ruckzug(error) | der(error) | Rückzug(correct) | gab(correct) | Jetzt(correct) | begannen(correct) | die(correct) | Skythen(correct) | die(correct) |
ect) | Verfolgung(correct) | um(correct)
1 0011: Perser(correct)
1 0012: schwärmten(correct) | das(correct) | Heer(correct) | von(correct) | allen(correct) | Seiten(correct) | und(correct) | bedrängten(correct) | es
(correct) | so(correct) | sehr(correct)
1 0013: Die(error) | Griechen(error) | daß(correct) | es(correct) | froh(correct) | sein(correct) | mußte(correct) | wieder(correct) | den(correct) | I
ster(correct) | zu(correct) | erreichen(correct) | Zum(correct) | Glück(correct)
1 0014: bei(correct) | der(correct)
1 0015: Brücke(correct)
l_0017: Fig(correct) | 8(correct) | Darius(correct) | mit(correct) | Dienern(correct) | uber(error) | ihm(correct) | Ormuzd(correct)
1 0018: Wandbild(correct) | im(correct) | Palast(correct) | zu(correct) | Persepolis(correct)
1 0019: fand(correct) | man(correct) | dort(correct) | die(correct)
1 0020: Brücken(correct) | noch(correct) | vor(correct)
```

fand man dort die

Brücken noch vor:

vertrieben. So blieben

lie Brücken erhalten

Ground Truth (Regions) Prediction (Lines) Prediction (Line coloring by word-level labels)

Rückzug der wichen zurück und ließen sich in keine offene Schlacht ein, sondern Skythen. Persern die Lebensmittel ausgingen und Darius den Befehl zum Rückzug der Rückzug gab. Jetzt begannen die Skythen die Verfolgung, um-Perser. schwärmten das Heer von allen Seiten und bedrängten es so sehr, Die Griechen daß es froh sein mußte, wieder den Ister zu erreichen. Zum Glück

bei der Brücke.



seine griechische Heimat.

Rückzug der wichen zurück und ließen sich in keine offene Schlacht ein, sondern Skythen. verwüsteten nur das Land und verschütteten die Brunnen, bis den Persern die Lebensmittel ausgingen und Darius den Befehl zum Rückzug der Rückzug gab. Jetzt begannen die Skythen die Verfolgung, um-Perser. schwärmten das Heer von allen Seiten und bedrängten es so sehr, Die Griechen daß es froh sein mußte, wieder den Ister zu erreichen. Zum Glück

fand man dort die Brücken noch vor: der Mann, dem Darius seine Rettung verlankte, war Histiäus, er Stadtherrscher

(Tyrann) von Milet einer der griechischen Städte an der Westküste Kleinasiens, Es waren nämlich Skythen an den Fluß gekommen und hatten den Griechen geraten, die Brücken abzubrechen. dann würde das ganze persische Heer zugrunde gehen; Miltiades, ein Griechenfürst vom Hellespont, hatte dem beigestimmt, da dann die Griechen Kleinasiens sich vom persischen Joche befreien könnten, aber Histiäus widersprach and meinte umgekehrt. ie Stadtherrscher hätten alle Ursache, die Rettung des persischen Königs und Heeres zu wünschen, da sie nur durch des

Königs Macht in ihrer Stellung geschützt würden. Ohne diesen Schutz hätten die Bürgerschaften ihrer Städte sie schon längst vertrieben. So blieben die Brücken erhalten und Darius war gerettet. Histiäus wurde belohnt, Miltiades floh in

Rückzug der wichen zurück und ließen sich in keine offene Schlacht ein, sondern Skythen. verwüsteten nur das Land und verschütteten die Brunnen, bis den Persern die Lebensmittel ausgingen und Darius den Befehl zum Rückzug der Rückzug gab. Jetzt begannen die Skythen die Verfolgung, um-Perser. schwärmten das Heer von allen Seiten und bedrängten es so sehr, Die Griechen daß es froh sein mußte, wieder den Ister zu erreichen. Zum Glück



Fig. 8. Darius mit Dienern (über ihm Ormuzd).

dankte, war Histiäus, der Stadtherrscher (Tyrann) von Milet. einer der griechischen Städte an der Westküste Kleinasiens. Es waren nämlichSkythen an den Fluß gekommen und hatten den Griechen geraten, die Brücken abzubrechen. dann würde das ganze persische Heer zugrunde gehen; Miltiades, ein Griechenfürst vom Hellespont, hatte dem beigestimmt. da dann die Griechen Kleinasiens sich vom persischen Joche befreien könnten, aber Histiäus widersprach und meinte umgekehrt. die Stadtherrscher hätten alle Ursache, die Rettung des persischen Königs und Heeres zu wünschen, da sie nur durch des Königs Macht in ihrer Stellung geschützt würden. Ohne diesen Schutz hätten die Bürgerschaften ihrer Städte sie schon längst vertrieben. So blieben die Brücken erhalten

fand man dort die

Brücken noch vor:

der Mann, dem Darius

seine Rettung ver-

und Darius war gerettet. Histiäus wurde belohnt, Miltiades floh in seine griechische Heimat.

und Darius war gerettet. Histiäus wurde belohnt, Miltiades floh in

Wandbild im Palast zu Persepolis.

seine griechische Heimat.

Ground Truth (Regions)

Prediction (Lines)

Prediction (Line coloring by word-level labels)

chtinier. Hier gewahren wir Gegensätze weit ausgeschnter Massen (Alpen), selartig aerstreater Berggrappen (Böhmerwald , Erzgebirge , Riesengebirge, edeten, Röhm. mahr. Höbentug, Karsi) und Verbindungen beider Pormen

Die fünfte Schichte (4000 bis 9000 Few), im bläulichen Tone und fer Voralpen, der Aposainon und den Puss jener meh beberen, mit ewigen Eis und Schnes bedeckten Gipfel, welche

die sochute Schichte (über 9000 Paus) bilden, und, weil am höchsten, ireh die dunkelste Farbe, das volle Schwarz, ausgezeichnet wurden. Nor in den lpen körsent diese Schiehte vor; die Karpathen erheben sieh nirgends au leher Höhe, ween sie gleich an einiges wenigen Punkten ihr nahe kommen Patra, Pogurascher Gebirge).

Nur wenige bewohnte Orte finden wir in den Rochthälers der Alpen. nsere Karte zeigt nur sehr wenige au, die (weil als kleine Orte mit schwarzen ntesen. Ein Vergleich der Karten 7 und 8 wird jeden entstehenden

Non am Schlusse der Erklürung der Karte angelangt, die so lange geworn, weil es galt, ein gans neues Bild verstehen zu lehren, ist der richtige late gegeben, jener Uebungen zu erwahnen, welche vor allem geeignet sind, empfangene, wohl aufgefanste Bild dem Gedächtnisse einzuprügen, und irch oftmalige Auwendung nachhaltig zu bewahren.

Dazu werden die Versnehe zur Lüsung folgendor Aufgaben führen: 1. Bestimmt die Numen aller Gewässer, die euch von der Karte von Uebrige findet sich dann feicht. ttel-Koroga ber sehon bekannt sind. Sochet aus den Kronländerharten die

-

chem Wald und Wiese vorherrschen, die Region des höheren Landen und der | Namen der Nebenfique, der kleineren Seen, kurr aller auf der Schichtenkan verkommenden Gewässer, Geht dabei in einer gewissen Ordnung vor, z. B.: Wählt merst die Donau mit füren Nebenfitzen und den in ihren Gebieton be findlieben Seen, daan die Elbe, die Oder, die Weichsel, den Po, die kleine

2. Bestimmt auf Shaliche Art die Lage der Bergheiten und Gipfel, die ihr orch Schruffen noch mitt berausgehabte, schliest die Hoot- und Alpen-skiege ein. Sie fant in sich die Stochgefel des Karutlandes, der Karpathen, Kenntnis dereh das Bertherrichen der feblenden Namen aus den Kronlände karten. Auf diese Art werdet ihr nach und nach von sehr vielen , sollist klein

irginseln Rechemehaft geben können. 3. Thut dasselbe ländermeise oder mich Plussgehieten, oder in irgend eine andern (immerkin aber in einer gewissen) Ordnung mit den Orten und bemerk dabei die Lage (an Flassen, Kanales, Scen, auf der Ebene, im Gebirge, a Eisenbahnen, am Meere u. s. f.), die Bevolkerungsklasse, die Verbindung n deen Orten durch Stolena, Rahnan.

4. Legt euch bei jedem Kroulando die Prage vor : Wie viel Sebiehten (der Hibenkarte) kommen darin ver? Ist eine darunter (und welche) überwiegen allen bezeichnet) von den Schnesgipfeln sorgnam neterschieden worden gegen die übrigen? Wie verhalt nich die Lage der grösseren Orte in Beziehung auf die Schiebten? In welcher Schichte liegen die meisten?

Wesn ihr solche Uebungen verständig unternehmet und eie auszufähr nicht ermüdet, werden die guten Früchte nicht lanen auf sich warten lasten. Ihr werlet auf diese Weise dahin kommen, die topische Geographie eines Lander soweit zie von der Kurte abgelesen werden kann, sell bet zu machen, und die wird each ungleich grösseren Nutzen bringen, als jedes noch so fleissige Aus rendiglernen aus einem Bache. Um dieses Ziel zu erreichen, muss eset Bestreben nuemt dahin geben, eine Kurte richtig lesen au lernen, das wrichem Wald und Wiese vorherrschen, die Region des biberen Landes und der i Namen der Nobestütste, der kleiseren Soon, kurr aller auf der Schichtenkarte

Die funfte Schiehte (4000 bis 9000 Fass), im biaulichen Toos und 2. Bestimmt auf thaliche Art die Lage der Bergketten und Gipfel, die ihr Elis und Schnee bedeckten Gipfel, welche

die nechate Schiehte (über 2000 Fun) bilden, und, meil um blebeten. 3. That daneibe landerweine oder nich Plumgebielen, oder in irgend oner

Nur wenige bewohnte Orte finden wir in den Hochthalers der Alpen. 4. Legt euch bei jedem Kronlande die Frage vor: Wie viel Schichten (der musten. Ein Vergleich der Karten 7 und 3 wird foden entstehenden auf die Behiehten? In welcher Behiehte liegen die meisten?

durch oftmalige Auwendurg nachhaltig zu bewahren.

I. Bestimmt die Namen aller Gewasser, die euch von der Karte von Uebrige findet sich dann leicht. Mittel-Europa ber sehra bekannt sind. Suchet nus den Kronfluderbarten die

Hochthaler, Hier gewahren wir Gegensates weit ausgelehnter Massen (Alpen), vorkommenden Gewässer, Gold dabei in einer gewissen Ordnung vor, z. B.: inselarig zerstreuter Berggruppen (Böhmerwald , Erzgebiege , Riesengebiege, Wahlt merst die Donau mit füren Neben Lissen und den in ihren Gebieten be-Sudeten, Bolim. mahr. Höhentung, Kurut) und Verbindungen beider Formen findlichen Seen, dann die Elbe, die Oder, die Weichsel, den Po, die kleinen

durch Sobruffen noch mehr berungehoben, schilenst die Hieb- und Alpen- nuf der Kurte von Mittel-Europa schien kennen femitet, und vervollständigt die gebirge ein. Sie fasst in sieht die Hochgiefel des Karellandes, der Karpathen, Kommins durch das Hernbernichen der fehlenden Namen aus den Kronländerder Voralpen, der Aponnison und den Puss jener noch höberen, mit ewigen Austen. Auf diese Art werdet ihr nach und nach von sehr vielen, seibst kleinen Berginseln Rechenschaft geben können,

durch die dunkelete Parbe, das volle Schwars, ausgeneichnet wurden. Nur in den undern (immerkin aber in einer gewissen) Ordnung mit den Orten und bemerket Alpen kommt diese Schichte vor; die Karpathen erheben sich nirgends au dabei die Lage (an Plance, Kanales, Seen, auf der Ebene, im Gebirge, an nolcher Höhe, wenn nie gleich an einigen wenigen Punkten ihr nahe kommen Eisenbahnen, am Meere u. n. f.), die Bevölkerungsklasse, die Verbindung mit andern Orten derch Striese, Robnen.

Unsere Karto zeigt nur sehr wenige au, die (weil als kleine Orte mit schwarzen Hobeskurte) kommen darin vor? Ist eine darunter (und welche) überwingend Nallen bezeichnet) von des Schneepipfeln sorgen unterschieden werden gegen die übrigen? Wie verbalt nich die Lage der gefoneren Orte in Besiehung

Wenn ihr solche Uebangen verständig unternehmet und eie auszuführen Non am Schlosse der Erklarung der Karte angelangt, die so lange gewor- nicht ermüdet, werden die guten Früchte nicht lange auf sich warten lassen. Ihr den, weil es gult, ein gann ne u.e. a Bild verstehen zu lehren, ist der richtige wertet auf diese Weise dahin kommen, die topische Geographie eines Landes, Plate gegeben, jener Uebungen zu erwähren, welche vor allem geeignet nind, soweit nie von der Karte abgelesen werden kenn, sol bat zu machen, und dies das ampfangene, wehl aufgefasste B.M dem Godachteines einzupragen, und wird euch ungleich grösseren Nutzen bringen, als jedes noch so fielesige Ausdi oftendige Ausendung nachkaltig au bewahren.

Dan werden die Vernebe zur Lieung felgender Aufgaben fideren:

Bestreben zuernt dahie geben, eine Karte riebtig lesen au lerenen, das wrichem Wald und Wiese vorharreshen, die Region des höberen Landes und der | Namen der Nebenflüsse, der Meineren Soon, kurn aller auf der Schichtenkarte

inselarity serstreuter Berggruppen (Böhmerwald, Erzgebiege, Rissengebirge, Wahlt merst die Donnu mit ihren Nebenfüssen und den in ihren Gebieten be-Sudeten, Holam. mahr. Hohameng, Karsi) und Verbindungen beider Pormen findlichen Seen, dann die Elbe, die Oder, die Weichsel, den Po, die kleinen

Die fanfte Schiehte (4000 bis 9000 Fass), im bifalichen Toos und 2. Bestimmt unf abnitiche Art die Luge der Berghetten und Gipfel, die ihr durch Schruffen noch mehr hernangehoben, schlienst die Hoeb- und Alpen- unf der Karte von Mittel-Buropa sehien kennen lerntet, und vervollständigt die gebiege ein. Sie fasst in sieh die Hochgiefel des Karellandes, der Karpathen, Konntein durch das Hernbernichen der fehlenden Namen aus den Kronländerder Voralpen, der Apennisen und den Foss jener noch höheren, mit ewigen karten. Auf diese Art werdet ihr nach und nach von sehr vielen, seihet bleisen

die nochute Schichte (über 2000 Pous) bilden, und, weil am biobsten, 3. Thut danselbe länderweise oder noch Flumgebieten, oder in irgend oner durch die dunkelste Farbe, das volle Schwarz, ausgezeichnet wurden. Nur in des undern (lumeralie aber in einer gewissen) Ordnung mit den Orten und bemerket Alpen kommt diese Schichte vor; die Karpathen erheben sich nirgends au dabei die Luge (an Flützen, Kanalen, Sons, auf der Ebene, im Gebirge, an solcher Hobe, wenn zie gleich an einigen wenigen Punkten ihr nabe kommen Eisenbalten, am Metre u. s. f.), die Bevölkerungsklasse, die Verbindung mit

Nur wenige bewohnte Orte finden wir in den Hochthalern der Alpen. 4. Legt euch bei jedem Kronlande die Frage vor: Wie viel Schichten (der massen, Ein Vergleich der Karten 7 und 8 wird jeden entstehenden auf die Schiehten? In welcher Schiehte liegen die weisten?

Nun am Schlume der Erklarung der Karte negelangt, die so lange gewor- nicht erntidet, werden die gaten Früchte nicht lange auf sielt warten lassen. Itst den, weil es galt, ein gant noue a Bild vereichen zu lehren, ist der richtige werdet auf diese Weise dahm kommen, die topische Geographie eines Landes, Plats gegeben, Jener Urbungen zu erwichten, welche vor allem geeignet nind, soweit zie von der Karte abgelesen werden kenn, sol bat zu machen, und dies das empfangene, wohl aufgefaute Sid dem Godantinine einupragen, und wird euch ungleich getweren Nutzen bringen, als jedes noch so fielesige Ausdurch oftmalige Anwendung nachhaltig au bewahren.

I. Bestimmt die Namen aller Gewasser, die euch von der Karte von Uchrige fiedet sich dann leicht. Mittel-Kuraga her sehon bekannt sind. Suchet uns den Kronlinderkarten die

Hochthaler, Hier gewahren wir Gegenstite weit ausgelehnter Masses (Alpen), vorkommenden Gewässer, Geht dabei in einer gewissen Ordnung vor. z. B.:

Berginseln Rechenschaft geben können,

andern Ortan derch Ströme Rahnan

Unsere Karte neigt nur sehr wenige au, die (weil als kleine Orte mit sehwarzen Hohenkarte) kommen darin vor? Ist eine darunter (ond welche) überwiegend Nullen bezeichnet) von den Schneegipfeln norgen unterschirden werden gegen die übrigen? Wie verhalt nich die Lage der grösseren Orte in Beziehung

Wesn ihr solche Uebangen verständig unternehmet und sie auszafül

wendiglernen aus einem Buche. Um dieses Ziel zu erreichen, muss ever Dazu werden die Versuche zur Lösung folgender Aufgaben führen: Bestreben zueret dahie gehen, eine Karte richtig lenen zu lernen, das

Ground Truth (Regions)

Prediction (Lines)

Prediction (Line coloring by word-level labels)

94

bas beseichtigte Lager bes Entschafteres an und ersocht einen glünzenden Sieg. Einige Tage später ergab sich auch die Feltung Belgrad. Unter den vielen Siegen, die der Prinz ersochten hat, ist der Verlägend der berühmteste. Noch heute singt man das damals entstandene Soldatenslied von "Prinz Engentus, dem edlen Kitter", der die "Brücke sieder die Vonzentus, den dem Kaiser "Stadt und Bestung Belgarad" zurückzueroben; noch heute erkönt dein Marisch der Sierreichsischen Reiterei die gleichfalls altertilmitiken Weite, and der den genen vorde.

neteten Jusolge bieses glängenden Sieges mußte die Afriei im Frieden zu wuser- Pa is a farounts (hr. Paligarowis) neuerdings ansgedehnte Gebiete an Die Freich vereich derteen, aminich das Vanat, einen Zell Wosnielens, der häter wieder verloren ging und erft in unsern Tagen wieder mit Osterreich vereinigt vurde, und iogaar große Leile der hentigen Königreiche Gerien und Rumänien. Vaha einem unhwoolene Leden, wie es Gerbien und Rumänien. And einem unhwoolene Leden, wie es Gerstelle von die Kunstellen zu teil geworden ist, starb der Prinz im Jahre 1736, tiessen die Freichschen zu teil geworden ist, starb der Prinz im Jahre 1736, tiessen ist die Konstellen von der Kossung zu Weisen (hurch den Pilikhauer Freinston) ein hertliches Reiterbenfunglererschieben lässen.

#### Bur Bieberholung:

Die pragmatische Sanktion. Aussterben des habsburgischen Mannsftammes. Fellischung der weiblichen Thronfolge durch Kaifer Karl VI. (Pragmatische Sanktion.) Anerkennung derfelben durch die Fürsten Europas.

Maria Cherelia. Ihre Erziehung. Bermählung mit dem herzoge Franz von Vothringen. Thronbesteigung Maria Theresias 1740. Bartenstein.

Augriff des Bönigs Friedrich II. von Preußen. Die übrigen Feinde Diterreichs (Bayern, Frankreich, Spanien). Schlechter Zustand des öfferreichischen Herreichischen Berdmangel.

Maria Therefia in Ungarn. Die ungarische Königskrönung. Der Reichstag zu Kreßburg. Siege der Österreicher über die Bayern. Rückzug der Franzosen aus Böhmen. Abtretung Schlesiens. Kaiserwahl und Krönung Franz I.

#### 21. Der fiebenjährige Rrieg.

Kaiserin Maria Theresia vor dem Kriege. Im helbenmittigen Rampfe un ibr gutes Recht hatte die große Kaiserin Maria Theresia oben größten Teil ihrer Erblande gegen die Here fast aller entopäischen Staaten bechamptet; nur ein einziges größeres Land hatte sie ab-

94

Anfolge biefes gläugenben Sieges mußte die Afrifei im Frieden gu westere Pa isaro wie stere Manifak des Panat, einen Zeil Bosnielens, der später wieder verloren ging und erft in unfern Tagen wieder mit Offerereich vereinigt wirde, und sogar große Zeile der heutigen Königreiche Verloren und Rumanien. Ich ach einem ruhmwollen Leden, wie es wester Gerbeit min Rumanien. Ich ach einem ruhmwollen Leden, wie es wester Gerbeit min Rumanien. Ich ach einem ruhmwollen Leden, wie es wester Gerbeit min Rumanien. Ich ach einem ruhmwollen Leden, wie es wester Gerbeit mit Bumanien. Ich ach einem ruhmwollen Leden, wie es wester Gerbeit mit Bumanien. Dach einem ruhmwollen Leden, wie es wester Gerbeit und Feiten der gewerben ist, staat der Prinz im Jahre 1736, tiefbetrauert von seinem damsbaren Wonarden, aber auch und dam gaugen öfterreichsten Bolfe. Aufer Franz sofie I. hat dem "Kitter ohne Artheit und Ladel" auf dem Plage vor der Hofburg zu Wind sein Vildhamer Kernson) ein herrliches Reiterbenstmal errichten lassen.

#### Rur Bieberholung:

Die pragmatische Sanktion. Anstierben des habsburgischen Mannsframmes. Keltebung der weiblichen Thronfolge durch Kaiser Karl VI. (Pragmatische Sanktion.) Anerkennung derselben durch die Kursten Europas.

Maria Cherefia. Ihre Erziehung. Bermählung mit bem Berzoge Thronbesteigung Maria Therefias 1740. Bartenftein.

Angriff des Königs Friedrich II. von Preußen. Die übrigen Beinde Sterreichs (Babern, Frantreich, Spanien). Schlechter Bustand bes öfterreichischen heeres. Geldmangel.

Maria Therefia in Ungarn. Die ungarifte Königstednung. Der Reichtung zu Bregburg. Siege ber Öfterreicher über bie Bauern. Midgug ber Frangolen ans Böhmen. Abtretung Schlesiens. Kaiferwahl und Kronung Frang I.

#### 21. Der fiebenjährige frieg.

Kauffe in Maria Cheresia vor dem Kriege. Im helbenmittigen Kampfe un ihr gutes Reicht hatte die große Kaifein Maria Dewesia ben größten Teil ihrer Erblande gegen die Seere fast aller europäischen Seaten beganptet; nur ein einziges größeres Land batte sie ab-

94

bas beseichtigte Loger bes Euthlahberes au und erscht einen glünzenden Sieg. Einige Tage häter ergab sich auch die Festung Belgrad. Unter den vielen Siegen, die der Prinz erschtlend des ist der Verlägenden des der bestillt der Belgrad der berühmteste. Noch hente singt man das damals entstundere Soldatenslied von "Prinz Eugentus, dem edlem Ritter", der die "Bride (tider die Vonum) sichgener" sies, und dem Kaiser "Stadt und Bestung Belgraad" zurückzuerobern; noch heute erkönt dein Marche der öfterreichsigen Reiterei die gleichfalls altertimusche Weise, nach der das Ees den general von der

#### Rur Bieberholung:

Die pragmatische Sanktion, Anssterben des habsburgischen Mannsftammes. Feltschung der weiblichen Thronfolge durch Kaifer Karl VI. (Pragmatische Santtion.) Anerkennung derselben durch die Kurtten Europas.

Alaria Cherefia. Ihre Erziehung. Bermafilung mit bem herzoge Braug bon Lothringen. Thronbesteigung Maria Therefias 1740. Bartenftein.

Angriff des Königs Friedrich II. von Preußen. Die übrigen Keinde Öfterreichs (Bayern, Frantreich, Spanien). Schlechter Zustand des öfterreichischen Beres. Geldmangel.

Maria Therefia in Ungarn. Die ungarifte Königströnung, Der Reichstag ju Prefburg. Siege der Öfterreicher über die Bauern. Ridgug ber Frangofen aus Böhmen. Abtretung Schlesiens, Kaiferwahl und Kronung Frang I.

#### 21. Der fiebenjährige Rrieg.

Kaiferin Aaria Cheresta vor dem Kriege. Im helbenmiltigen Rampfe un ihr gutes Recht hatte die große Kaiferin Maria Theresta den größten Teil ihrer Erblande gegen die Geere fast aller entopälischen Staaten behanptet; nur ein einziges größeres Land hatte sie ab-

## **Tokenization and Model Training**

#### **Tokenization**

- Use a pre-trained transformer tokenizer (e.g., DistilBERT for German).
- Each line is represented as a sequence of words and their assigned labels.
- For tokenization, each word is split into one or more sub-tokens as required by the tokenizer.
- Only the first token of each word is assigned the word's label; all subsequent sub-tokens are assigned a special label (-100) so they are ignored during model training.
- This approach maintains word-level supervision even when words are split into subtokens, allowing the model to learn from the original labeled data.

## **Tokenization and Model Training**

#### Model Training:

- Fine-tune a pre-trained transformer (e.g., DistilBERT) for token classification (labels: correct/error)
- Input: Tokenized sequences of words per line, with aligned labels.
- Output: The model predicts the label (correct/error) for each word in the input sequence.
- Training parameters:
  - Learning rate: 2e-5
  - Batch size: 8 (train & eval)
  - o Epochs: 10
  - Weight decay: 0.01

## **Tokenization and Model Training**

#### **Model Training:**

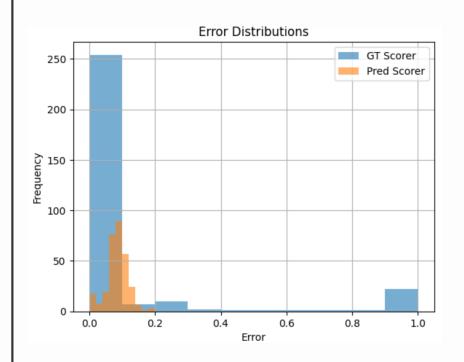
#### Class Imbalance Handling:

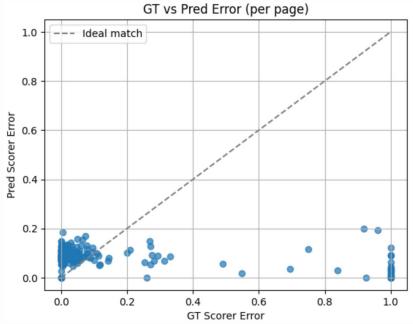
- Compute class weights inversely proportional to the frequency of each class (correct/error).
- Use a custom training loop that applies these weights in the loss function to handle class imbalance and improve error detection.

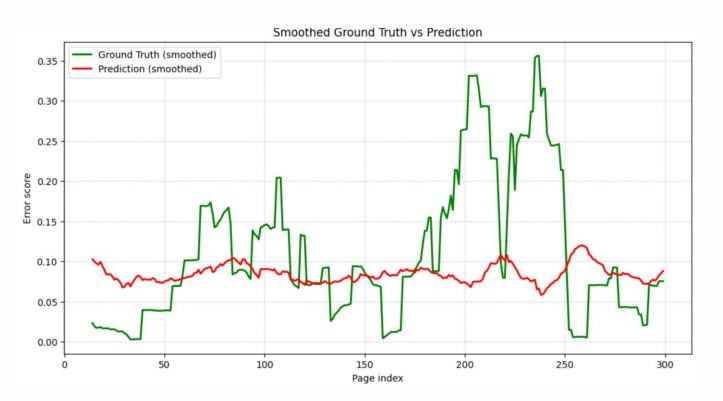
#### Custom Trainer

- Use a custom Trainer that applies weighted cross-entropy loss, giving more importance to rare classes (like errors).
- During loss calculation, we ignore all sub-tokens (i.e., tokens with label -100), so only the main word tokens contribute to model training.

+	++	+
·	GT Scorer   	Pred Scorer
Count	300   	300.0000
Mean	0.1078	
Median	0.0048	0.0859
Stdev	0.2685	·
Min	0	0.0000
Max	1	0.1991
MAE	0.1078	•
Correlation	-0.3678	-
GT vs Pred RMSE	0.2827	-







t	cop-30 GT pages by relative error		top-30 Predicted pages by relativ
Rank	Page	RelErr	Rank Page
	1727264816 0000001	1 0000	1 1741701600 00000004
J	1737364816_00000001	1.0000 1.0000	1 1741701600_00000024
2	1737377780_00000004		2 174170569X_00000012
3	1737379171_00000132	1.0000	3 1744303657_00000355
4	1737379627_00000003	1.0000	4 1740392299_00000023
5	1737386917_00000003	1.0000	5 1744303657_00000369
6	1737389673_00000080	1.0000	6 1740404564_00000112
7	1737393700_00000004	1.0000	7 1741193575_00000027
8	1738471039_00000010	1.0000	8 1737386917_00000328
9	173857282X_00000390	1.0000	9 1739039688_00000203
10	1738584399_00000150	1.0000	10 1740405188_00000036
11	1738584852_00000167	1.0000	11 1740379381_00000113
12	1739039688_00000224	1.0000	12 1740399145_00000128
13	1740257367_00000080	1.0000	13 1740400666_00000080
14	1740257367_00000087	1.0000	14 1740395999_00000052
15	1740384288_00000002	1.0000	15 1737381346_00000170
16	1740388542_00000002	1.0000	16 1740392299_00000018
17	1740391969_00000066	1.0000	17 1741701600_00000061
18	1740391969_00000080	1.0000	18 1740405188_00000011
19	1740399145_00000435	1.0000	19 173736378X_00000120
20	1740399145_00000436	1.0000	20 1741704316_00000001
21	1740399404_00000002	1.0000	21 1737381346_00000249
22	1740399404_00000010	1.0000	22 1740391551_00000049
23		1.0000	23 174429397X_00000011
24		1.0000	24 1739037960 00000033
25	1740402723 00000010	1.0000	25 1737359839 00000247
26	1740404564 00000001	1.0000	26 174170569X 00000008
27	1740406060_00000003	1.0000	27 1744303878_00000108
28	1740826949_00000062	1.0000	28 1740388542 00000038
29	1741190975_00000026	1.0000	29 1744303878_00000068
30	1741190975_00000043	1.0000	30 1740404947 00000238
		2.0000	20 27 . 10 . 10 . 17 . 1 . 200000 = 1

```
--- page-level error statistics --- GT pages : \mu = 0.2443 Pred pages : \mu = 0.0811 RMSE : 0.4507
```