EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Universitätsbibliothek Tübingen

OCR-BW

Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen

Transkribus-Workshop

Advanced Funktionen

Dorothee Huff

Übersicht

1. Sample Sets
2. Modelltraining
    1. HTR+
    2. PyLaia
3. P2Pala
4. Text2Image
5. Tabellen

© 09.06.2021 Universität Tübingen

Erstellung eines Sample Sets

- Zweck: Berechnung der wahrscheinlichen Performanz von Modellen anhand einer Stichprobe
- Tools → Compute Accuracy… → Compare Samples…
- Auswahl der gewünschten Dokumente → +Add to Sample Set
- Auswahl der gewünschten Zeilenauswahl
- Festlegung eines Schwellenwerts für die Zeilenlänge, damit keine zu kurzen Zeilen aufgenommen werden
- Erstellung von GT für das Sample Set, die zum Abgleich dient

© 09.06.2021 Universität Tübingen

Vergleich von Modellen auf einem Sample Set

- Auswahl eines oder mehrerer Modelle → Start der Text Recognition auf dem Sample Set (dies ist gebührenfrei)
- unter dem Reiter „Samples“ im Fenster „Compare Samples“ kann nun aus der „Sample Collection“ das gewünschte Sample Set ausgewählt werden, um die Berechnung zu starten
- Unter „Select hypothese by toolname“ Auswahl des Models → Compute → Anzeige der Ergebnisse
- Die Ergebnisse bleiben in einer Liste gespeichert und können auch später erneut abgerufen werden

© 09.06.2021 Universität Tübingen

Modelltraining

- Modelltraining ist aktuell mit zwei Engines möglich
  - HTR+
  - PyLaia

- Grundlegende Optionen
  - Auswahl des Trainingsmaterials aus den Dokumenten einer Collection
  - Auswahl des Trainingsmaterials aus der „HTR Model Data“ (so lassen sich schnell einzelne Parameter ändern)

  - Eigene Aufteilung des Materials in ein Training Set und ein Validation Set
  - Automatische Aufteilung (2%, 5% oder 10%)

  - Auswahl der Transkriptionsversion

© 09.06.2021 Universität Tübingen

Optionen beim Training mit HTR+

- Anzahl der Epochen
- Hinzuschaltung eines Base Models
- Auslassung von Zeilen, die die Tags „Gap“ und „Unclear“ beinhalten
- Umkehr des Texts (relevant z.B. für Schriften, die von rechts nach links verlaufen)

© 09.06.2021 Universität Tübingen

Optionen beim Training mit PyLaia

- zusätzlich zu den Einstellungsmöglichkeiten beim Training mit HTR+ können weitere Parameter angepasst werden

- wichtig: grundsätzlich kann zwar ein Base Model genutzt werde, jedoch muss das Trainingsmaterial exakt den gleichen Zeichensatz enthalten, da das Ergebnis sonst fehlerhaft ist

- Early Stopping: wenn nach der eingestellten Höchstzahl keine Verbesserung im Trainingsprozess auftritt, wird das Training beendet → dies kann manchmal besonders bei kleinen Modellen zu früh geschehen → eine Erhöhung des Werts kann sinnvoll sein

© 09.06.2021 Universität Tübingen

Optionen beim Training mit PyLaia

- Learning Rate: gibt an, wie schnell während des Trainingsprozesses von einer Epoche zur nächsten gewechselt wird, also wie schnell das Training verläuft

- Image type: Wahl zwischen dem Original Bild und einer komprimierten Version → letzteres ist sinnvoll, wenn die Vorverarbeitung der Bilder beim Training zu lange dauert

© 09.06.2021 Universität Tübingen

Advanced Parameters beim Training mit PyLaia

- Preprocessing:
  - beim Training von gedrucktem Material kann es das Ergebnis verbessern, wenn die Haken bei „Deslant“ und „Deslope“ entfernt werden

- Model:
  - Parameter sollten nicht verändert werden

- Training:
  - Batch size: Anzahl der Seiten, die gleichzeitig verarbeitet werden → z.B. Änderung auf einen Wert von 12

© 09.06.2021 Universität Tübingen

P2PaLa

- Zweck: Training eines Strukturmodells für eine verbesserte automatische Layouterkennung (das Modell lernt anhand von Größe und Platzierung auf der Seite, wo sich welche Textregion befindet)

- Other Tools → P2PaLa…

- Optionen
  - Rectify regions: alle erkannten Regionen werden in Rechtecke umgewandelt
  - Min area: Entfernung von kleinen Textregionen

© 09.06.2021 Universität Tübingen

P2PaLa - Training

- Voraussetzung: Strukturtagging der Textregionen
- mindestens 50-100 Seiten
- Auswahl der gewünschten Strukturtypen (nicht mehr als 5)
- Auswahl, ob Textregionen und/oder Baselines trainiert werden sollen

© 09.06.2021 Universität Tübingen

Nachnutzung von vorhandenen Transkriptionen - manuell

- Eine Transkription für eine Seite, die zeilengenau vorliegt, kann im ganzen kopiert und durch einen Klick in die erste Zeile der entsprechenden Seite in Transkribus eingefügt werden

- Nachkorrekturen sind natürlich möglich, aber es ist sinnvoll, das Layout vorher zu überprüfen

© 09.06.2021 Universität Tübingen

Nachnutzung von vorhandenen Transkriptionen - automatisch

- Voraussetzung: die Transkription liegt in einzelnen XML-Dateien in einem Ordner vor (pro Seite eine eigene Datei)

- Hauptmenü → Document → Sync transcriptions with doc… → Auswahl und Upload des entsprechenden Ordners → Überprüfung der Zuordnung (eine vorherige Layoutanalyse ist nicht notwendig)

- Praktisch z.B. wenn eine Transkription aus Transkribus exportiert und wieder importiert werden soll

© 09.06.2021 Universität Tübingen

Nachnutzung von vorhandenen Transkriptionen – automatisch mit Text2Image

- Zweck: große Mengen vorhandener Transkriptionen automatisch einspeisen

- Voraussetzung: die Transkription liegt in einzelnen Seiten als TXT-Dateien in einem Ordner vor (am besten sind die einzelnen Dateien entsprechend der Seiten benannt)

- Upload der TXT-Dateien über: Hauptmenü → Document → Sync local text files with doc… → Auswahl und Upload des entsprechenden Ordners → Überprüfung der Zuordnung (keine Layoutanalyse)

- Other Tools → Text2Image…→ Auswahl eines Basemodels →Run

© 09.06.2021 Universität Tübingen

Tabellen

- Tabelle erstellen: Canvas Menu → Add other item… → Table → Unterteilung in Zeilen und Spalten über die Schneidwerkzeuge

- Übertragung der erzeugten Tabelle auf weitere Seiten: Canvas Menu → copy selected regions or tables…

© 09.06.2021 Universität Tübingen

Informationen

Registrierung/Download Transkribus: https://readcoop.eu/de/transkribus/download/
Allgemeine Informationen: https://readcoop.eu/

Anleitungen und Tutorials: https://readcoop.eu/transkribus/resources/

Blog vom Universitätsarchiv Greifswald, der eine gute Ergänzung zu den von Transkribus zur Verfügung gestellten Anleitungen darstellt, und einzelne Aspekte näher beleuchtet:
https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/de/

Videotutorials von Annemieke Romein

- https://www.youtube.com/watch?v=5YCfaFNMol4
- https://www.youtube.com/watch?v=yxLyzRZaff8
- https://www.youtube.com/watch?v=axYRvhgcVF4

https://ocr-bw.bib.uni-mannheim.de/projektuebersicht/
https://uni-tuebingen.de/de/179298

© 09.06.2021 Universität Tübingen

Danke.

Kontakt: Dorothee Huff

Universitätsbibliothek Tübingen
Wilhelmstraße 32, 72074 Tübingen
Telefon: +49 7071 29-72852
Dorothee.huff@uni-tuebingen.de

© 09.06.2021 Universität Tübingen