READ
Recognition and Enrichment of Archival Documents

Testprojekte in Transkribus - für Archive und Bibliotheken

Dies ist eine kurze Einführung zur Erstellung eines Test- oder Pilotprojekts in Transkribus. Durch die Arbeit mit der Plattform erhalten Sie Zugang zu der neuesten Technik im Bereich der Texterkennung (Handschriften sowohl als auch gedruckte Schriften). Transkribus ermöglicht es, diese Technologie auf Ihre Dokumente anzuwenden.

Laden Sie den Transkribus Expert Client herunter oder stellen Sie sicher, dass Sie die neueste Version verwenden:

- https://transkribus.eu/

Besuchen Sie das Transkribus Wiki für mehr Informationen und weitere How to Guides:

- https://transkribus.eu/wiki/

Transkribus und die zugrundeliegende Technologie werden durch die folgenden Projekten und Plattformen verfügbar gemacht:

- https://read.transkribus.eu/
- https://transcriptorium.eu/
- https://github.com/transkribus/

Kontakt

- Das Transkribus Team: email@transkribus.eu

Inhalt

Das READ Projekt wird durch das Horizon 2020 Forschungs- und Innovationsprogramm im Rahmen des Fördervertrags Nr. 674943 finanziert.

Einführung

- Als Archiv oder Bibliothek sind Sie wahrscheinlich für eine (große) digitale Kollektion von handgeschriebenen Dokumenten verantwortlich.
- Sie möchten diese Dokumente verfügbar machen und Ihren Nutzern ermöglichen, in den Dokumenten auf neue Art und Weise zu suchen. Dies geschieht mit Volltext- (fulltext), mit Stichwortsuche (keyword spotting) oder mit Namenserkennung (named entity recognition).
- Sie sind interessiert daran, die Handschriftenerkennung (HTR) zu nutzen, um automatisch Transkriptionen der Dokumente in Ihrer Kollektion zu generieren.
- Sie möchten wissen, wie die HTR Technologie funktioniert und wie effizient sie sein kann.

Ein Testprojekt entwerfen

- Es ist einfach mit Transkribus ein Testprojekt zu erstellen, dass es Ihnen ermöglicht:
    - ein HTR Modell zu trainieren, das speziell Ihre Dokumente erkennt.
    - die Genauigkeit der Texterkennung auf wissenschaftliche Art und Weise zu evaluieren.
    - die Ergebnisse auf die ganze Kollektion hochzurechnen.
    - eine Einschätzung der Zeit und Ressourcen, die nötig sein werden, um Ihre gesamte Kollektion zu bearbeiten.
- Sie können Ihr Testprojekt jetzt gleich starten. Die Funktionen und Workflows sind verfügbar und können über das Transkribus Expert Interface gestartet werden.

HTR trainieren

- HTR Maschinen basieren auf maschinell lernenden Algorithmen, genauer gesagt auf überwachtem maschinellen Lernen. Das heißt, dass der HTR Maschine richtige Beispiele von transkribierten Dokumenten gezeigt werden müssen, sodass die Maschine die Muster der Zeichen und Wörter versteht.
- Generell kann man sagen, dass die Ergebnisse umso besser sind, umso mehr Trainingsdaten zur Verfügung stehen. Das gilt vor allem für Kollektionen, die viele verschiedene Autoren oder Schreibstile beinhalten.
- Im Rahmen des READ Projekts arbeiten Forschungsgruppen unter anderem an der Entwicklung eines einheitlichen Modells, dass die Summe aller Trainingsdaten integriert. Wenn schon Trainingsdaten verwendet wurden, die Ihren Dokumenten ähneln, erleichtert dies die Erstellung einer HTR für Ihre Kollektion.

Auswahl des Datensatzes

- Um das Testprojekt laufen zu lassen, braucht man einen “zuverlässigen” Datensatz. Dieser sollte ein repräsentatives Beispiel für die Dokumente in Ihrer Kollektion sein. Wenn man einen solchen Datensatz auswählt, sollte es möglich sein, die Ergebnisse auf die gesamte Kollektion hochzurechnen.
- Der Datensatz dient als Trainingsmaterial für die HTR und wird auch dazu verwendet, das Ergebnis zu evaluieren.
- Wir empfehlen die Dokumente aus Ihrer Kollektion dazu “willkürlich” auszusuchen. Der objektivste Weg ist die Dokumente automatisch mit Hilfe einer Datenbank auszuwählen oder mit einem einfachen Kriterium wie jedes zehnte oder jedes zwanzigste usw. Dokument vorzugehen.

Seitenanzahl

- Als Faustregel kann man sagen, dass 20 000 Wörter (ca. 100 Seiten) für ein Training schon genug sein können, sofern es sich um eine einfache Kollektion handelt, zum Beispiel ein Tagebuch oder Briefe, geschrieben von ein und derselben Person.
- Wenn Ihre Dokumente Schriften von verschiedenen Autoren enthalten und/oder mehrere Jahrzehnte oder Jahrhunderte umfassen, empfehlen wir einen Datensatz von mehreren Hundert Seiten auszuwählen.
- Nichtsdestotrotz können erste Tests schon mit kleineren Datenmengen gemacht werden und die Trainingsdaten dann, abhängig von dem Ergebnis, das die HTR erzielt hat, erhöht werden.

Der Datensatz wird zu Ground Truth

- Bevor der Datensatz als Training Set verwendet wird, muss er so aufbereitet werden, dass die HTR Maschine damit arbeiten kann.
- Diese fertig präparierten Daten nennen sich “ground truth” oder “Referenzdaten”, da sie die Basis für alle folgenden Vorgänge bilden.
- Bei der Vorbereitung von “ground truth” Daten gibt es zwei Hauptschritte:
    - **Segmentierung**
      Die Linien des Transkriptionstextes müssen mit den Linien im Bild/Scan verbunden werden. Um dies zu erreichen muss jedes “Image” in “text regions”, “lines” und “baselines” unterteilt werden. Mehr Details zur Durchführung finden Sie hier: Transkribieren mit Transkribus - Einführung.
    - **Transkription**
      Es ist eine korrekte Transkription der Texte im Dokument erforderlich. Der Transkriptionstext sollte so nah wie möglich am Ausgangstext sein, jeder Buchstabe im Dokument sollte durch das dementsprechende Zeichen in der Transkription präsent sein.
- Für moderne Dokumente, zum Beispiel 18. Jahrhundert aufwärts, ist die Transkription gewöhnlicherweise unkompliziert. In Dokumenten von früheren Zeiten können unübliche Zeichen und Abkürzungen zur Herausforderung werden. Um diese Arbeit zu erleichtern bietet Transkribus ein Tagging System an.

Den Datensatz transkribieren

- Sobald Sie die Seiten für den Datensatz ausgewählt haben, können diese ins Transkribus geladen werden und der Transkriptionsprozess kann beginnen.
- **Anmerkung:** alle Dokumente in Transkribus sind privat, keine anderen User haben darauf Zugriff.
- Es gibt zwei Wege die Transkription in Transkribus auszuführen:
    - **Option 1:** Sie transkribieren
      Sie oder Ihre Mitarbeiter oder Kollegen transkribieren den Text im Transkribus. In diesem Fall müssen Sie sich mit Transkribus vertraut machen, das dauert durchschnittlich zwei bis drei Stunden und sollte zum Großteil in Form von “learning by doing” möglich sein. Detaillierte Erklärungen zum Segmentieren und Transkribieren finden Sie in dieser Anleitung: Transkribieren mit Transkribus - Einführung.
    - **Option 2:** Wir transkribieren
      Wenn bereits eine Transkription existiert, bieten wir ein “Text2Image matching tool” an, welches die Transkriptionen den digitalisierten Bildern automatisch zuordnen

kann. Außerdem gibt es noch die Möglichkeit, dass Studenten oder externe Serviceanbieter die Transkription übernehmen. Sie haben Erfahrung in der Arbeit mit Transkribus und europäischen Sprachen. Der Preis hängt von der geforderten Genauigkeit der Transkription und Schwierigkeit der Handschrift ab.

HTR Training

- Sobald der Datensatz komplett transkribiert ist und als "ground truth" bezeichnet werden kann, kann das Training der HTR Maschine beginnen.
- Das Training läuft offline innerhalb der Transkribus Plattform. Meist dauert es wenige Wochen bis das neue oder überarbeitete HTR Modell verfügbar ist.
- Sobald das Training abgeschlossen ist, erhalten Sie eine Benachrichtigung und können dann das Modell nutzen um automatische Transkriptionen für den Rest Ihrer Dokumente zu produzieren und um in den Dokumenten zu suchen.

Evaluation der Ergebnisse

- Ein kleiner Teil des Datensatzes wird als "test set" für die Evaluation ausgespart. Dieser wird nicht für das Training genutzt.
- Da das Modell diese Seiten vor der Evaluation noch nicht „gesehen" hat, kann verlässlich die Präzision des Modells getestet werden.
- Zu diesem Zweck wurde in Transkribus ein Modul eingebaut, das die "Character Error Rate" und "Word Error Rate" errechnet, beides anerkannte Metriken in der Informatik.
    - **Anmerkung:** Sobald ein HTR Modell trainiert und in Transkribus verfügbar ist, können Sie es auf jede beliebige Seite anwenden, auch auf Seiten, die nicht Teil des ursprünglichen Datensatzes waren. Die Präzision des Modells kann auch für einzelne Seiten gemessen werden.
- Aktuelle Resultate in der Informatik zeigen, dass eine Character Error Rate utner 10% und eine Word Error Rate unter 20% den letzten Stand der Technik repräsentieren. Unter Laborbedingungen konnten sogar noch bessere Resultate erzielt werden.

Ihr Vorteil aus der Arbeit mit HTR

- Die HTR Maschine produziert eine automatische Transkription Ihrer Dokumente.
- Außerdem produziert sie "confidence matrices" auf Zeichen- und/oder auf Wortbasis. So werden die Optionen, die die HTR in Betracht gezogen hat, gespeichert.
- Basierend auf diesen "confidences" funktionieren zwei weitere Features:
    - **Keyword Spotting (KWS)**
      KWS ist ein Funktion mit der man nach bestimmten Wörtern in Dokumenten suchen kann, diese werden meist auch gefunden, wenn die Transkription nicht optimal ist. KWS sucht nach den Wörtern mit allen Optionen der HTR. Das erhöht die Chancen, das korrekte Wort oder die korrekten Wörter zu finden, auch wenn sie nicht den höchsten Wahrscheinlichkeitswert haben.
    - **Computergestützte Transkription**
      "Confidence matrices" sind auch bei der Korrektur von transkribierten Seiten nützlich. Alternative Wörter können gezeigt werden oder das Interface kann naheliegende Wörter vorschlagen, basierend auf vorhergegangenen Eingaben. Eine Demo Version von computergestützter Transkription ist auf der Website vom tranScriptorium Projekt verfügbar.

Strukturelle Daten

- Viele Archivdokumente sind strukturell organisiert, das heißt mit Tabellen oder mit Formularen mit sich wiederholenden Elementen. Diese zusätzliche Information kann natürlich genutzt werden um den Wert eines automatischen Transkriptionsprozesses zu erhöhen.
- Gute Resultate der HTR sind eine wichtige Voraussetzung für jede Art von struktureller Anreicherung. Außerdem sind strukturelle Informationen weniger normiert, als Schreibstile und müssen deshalb manuell bearbeitet werden, zum Beispiel mit der Anwendung regelbasierter Systeme. Dies erfordert Expertenwissen und die Einbindung von Informatikern.
- Wir empfehlen, dass Sie die ersten Tests nur auf die Zuverlässigkeit der HTR fokussieren und danach die strukturellen Informationen hinzufügen.

Nächste Schritte

- Wenn Sie an einem Testprojekt interessiert sind, empfehlen wir, vor Beginn mit uns in Kontakt zu treten, um die Eckpunkte Ihres Projekts zu besprechen. (email@transkribus.eu).
- Des Weiteren gibt es die Möglichkeit in Form eines “Memorandum of Understanding (MOU)” Teil des READ Projekts zu werden. Damit können Sie hinter die Kulissen schauen und bekommen Informationen aus erster Hand. Auf der READ Project Website finden Sie eine Liste der Bibliotheken und Archive, die bereits ein MOU unterzeichnet haben.

Danksagung

Wir möchten den vielen Nutzern danken, die mit ihrem Feedback zur Verbesserung der Transkribussoftware beigetragen haben.

Transkribus wird der Öffentlichkeit im Rahmen des H2020e Infrastrukturprojekts READ (Recognition and Enrichment of Archival Documents) zugänglich gemacht, das von der Europäsichen Kommission finanziert wird.