Der Goldstandard-Datensatz von JobAds

Der Goldstandard-Datensatz ist ein manuell korrigierter Datensatz aus 14.873 transkribierten historischen Jobinseratssegmenten, der mithilfe von Studienassistent:innen zwischen 2024 und 2025 im Rahmen des JobAds-Projekts an der Universität Graz erarbeitet wurde. Er wurde mit dem Ziel erstellt, eine hochwertige Datengrundlage für die Evaluierung von OCR (Optical Character Recognition)-Modellen und anschließende Textanalyse-Experimente mit historischen Stelleninseraten zu schaffen.

Der Goldstandard basiert auf dem ebenfalls innerhalb des JobAds-Projekts erstellten Ground Truth-Datensatz der Seitensegmente. Dieser enthält 14.985 manuell annotierte Segmente, welche einzelne auf historischen Zeitungsseiten identifizierte Stelleninserate darstellen.
Die Ground Truth wurde als Teil des CHR2024-Konferenzpapiers von Venglarova, Adam, Mölzer et al. (2024) veröffentlicht und ist im CSV Format auf der GitHub-Plattform https://github.com/JobAds-FWFProject/Ground-Truth-CHR2024 verfügbar. Der Datensatz kann unter der Lizenz CC BY 4.0 verwendet werden.

Während im Ground Truth-Datensatz die Kategorie und die Position des Stelleninserats auf der jeweiligen Zeitungsseite ermittelt wurde, liefert der Goldstandard eine korrigierte Transkription des in den annotierten Segmenten enthaltenen Textes. Die zwei Datensätze ergänzen einander also: die Ground Truth gibt Auskunft über die Position eines Inserats sowie dessen Kategorie (Jobangebot, Jobsuche, Angebot einer Dienstleistung, Überschrift, Vermittlung oder Indikator), und der Goldstandard enthält den Text des Inserats.

Die unbearbeiteten OCR-Transkriptionen, aus denen der Goldstandard erarbeitet wurde, wurden mithilfe des frak2021 OCR-Modells erstellt (Universitätsbibliothek Mannheim, 2021). Diese OCR-Rohdaten wurden mittels der Transkribus-Plattform (Kahle et al., 2017) manuell überprüft und korrigiert. Bei der Transkription wurde auf eine möglichst originalgetreue Wiedergabe der gedruckten Texte Wert gelegt, deshalb wurden diese nicht an die aktuelle Rechtschreibung oder aktuelle drucktechnische Standards angepasst.

Demzufolge blieben historische und typographische Eigenheiten des Quellenmaterials erhalten. Zum Beispiel wurde das historische lange *s* (*ſ*) beibehalten anstatt es in das moderne *s* umzuwandeln. Ebenso wurden die ursprüngliche Silbentrennung, die Abstände sowie Rechtschreibung und Druckfehler beibehalten. Dieser Ansatz soll sicherstellen, dass der Goldstandard so nah wie möglich an der historischen Quelle bleibt und als geeignete Referenz für die Bewertung von OCR-Systemen in Bezug auf historische Dokumente herangezogen werden kann**.** Annotatoren wurden eingesetzt, um OCR-Fehler auszubessern oder auch die Wortreihenfolge einzelner Textzeilen richtigzustellen, wo diese vom OCR-System nicht richtig erkannt worden war. Völlig unleserliche Wörter wurden mit ‚####' dargestellt.

Der Goldstandard wurde, wie die Ground Truth auch, nur in einem Annotations- bzw. Korrekturdurchgang erstellt und konnte aufgrund beschränkter Mitarbeiterkapazitäten nicht einer nochmaligen systematischen Prüfung unterzogen werden. Daher kann der Datensatz vereinzelt fehlerhafte manuelle Korrekturen bzw. übersehene OCR-Fehler enthalten. Die mögliche Anzahl derartiger Fehler kann zum jetzigen Zeitpunkt nicht quantifiziert werden, darum sollte der Goldstandard nicht als absolut fehlerfreie Abschrift betrachtet werden, sondern vielmehr als ein mit größter Sorgfalt händisch bearbeiteter Referenzdatensatz.

Datenformat

Der Goldstandard wurde im CSV-Format erstellt. Jede Reihe entspricht einem der 14.873 annotierten Stellenanzeigensegmenten und enthält Metadaten zur Zeitungsquelle mit Seitenangabe, zu Positionierung und Anzeigenkategorie, das ursprüngliche OCR-Ergebnis and die manuelle Korrektur davon. Der Datensatz umfasst etwas weniger als die in der Ground Truth enthaltenen 14.985 Segmente, da einige davon im Ausgangsmaterial unleserlich waren und nicht transkribiert werden konnten – diese wurden daher nicht in den Goldstandard übernommen.

Der Datensatz enthält folgende Spalten:

Spalte	Beschreibung
`pub`	Kennzeichnung der Zeitung (z.B.: `gtb` für *Grazer Tagblatt*).
`pub_date`	Erscheinungsdatum der Zeitung (z.B.: 02031860 für 2.3.1860).
`page_num`	Nummer der Zeitungsseite.
`page_id`	Eindeutige Identifikation der Zeitungsseite in Kombination mit Zeitungsname und Erscheinungsdatum.
`uuid`	Eindeutige Kennzeichnung des einzelnen Anzeigensegments.
`label`	Zugeordnete Anzeigenkategorie in der Ground Truth: Stellenangebot, Stellengesuch, Angebot einer Dienstleistung, Überschrift, Vermittlung oder Indikator.
`x`	Horizontale Koordinate des Begrenzungsrahmens der Anzeige auf der Seite, von links oben gemessen.
`y`	Vertikale Koordinate des Begrenzungsrahmens der Anzeige auf der Seite, von links oben gemessen.
`width`	Breite des Begrenzungsrahmens der Anzeige.
`height`	Höhe des Begrenzungsrahmens der Anzeige.
`ocr`	Vom OCR-Modell frak2021 erstellte OCR-Transkriptions-Rohdaten, inklusive Zeilenumbrüche.

`ocr_without_lb`	OCR-Transkriptions-Rohdaten ohne Zeilenumbrüche.
`text`	Manuell korrigierte Goldstandard-Transkription unter Beibehaltung der Zeilenumbrüche und der Charakteristika des gedruckten Originals.
`text_without_lb`	Manuell korrigierte Goldstandard-Transkription ohne Zeilenumbrüche.
`iiif`	IIIF-Link zum Bild des jeweiligen Inserats, das einen unmittelbaren Vergleich der Transkription mit dem original Zeitungsinserat ermöglicht.

Anhand der Koordinatenangaben (`x`,`y`,`width` und `height`) definiert sich der rechteckige Begrenzungsrahmen eines jeden Inseratssegments auf der entsprechenden Zeitungsseite. Die Felder `ocr` und `text` beinhalten Zeilenumbrüche, während ihre Entsprechungen `_without_lb` die Texte ohne Zeilenumbrüche bereitstellen, was für automatisierte Textaufbereitung und OCR-Auswertungen nützlich sein kann.

Die Felder `ocr` und `text` ermöglichen den direkten Vergleich zwischen dem automatischen OCR-Ergebnis und der manuellen Goldstandard-Transkription. Im Feld `iiif` erhält man direkten Zugang zum Bild der Anzeige, was die manuelle Überprüfung und weitere Forschung auf Basis des Originalmaterials erleichtert.

Referenzen

Kahle, P. et al. (2017). Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). pp. 19–24. 10.1109/ICDAR.2017.307.

Universitätsbibliothek Mannheim (2021). Frak2021.https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/frak2021/tessdata_best/frak2021-0.905.traineddata (Zugriff vom 20. November 2023).

Venglarova, K., Adam, R., Mölzer, W., et al. (2024). Who Advertises in Newspapers? Data Criticism in Mining Historical Job Ads. In *Proceedings of the Computational Humanities Research Conference 2024*. Computational Humanities Research 2024. Aarhus, Denmark: CEUR-WS, pp. 788–801.ceur-ws.org/Vol-3834/.