Praktische Probleme des OCR-Trainings mit synthetischen Daten

Kay-Michael Würzner
wuerzner@bbaw.de

Gemeinsame Arbeit mit Ehud Alexander Avner und Matthias Boenig

OCR-Entwicklerworkshop an der Berlin-Brandenburgischen Akademie der Wissenschaften
28. September 2017

Übersicht

- Einleitung
  - Was ist synthetisches Training?
  - Wozu braucht man das?
- Technische Aspekte
  - Font-Rendering
  - Schriftarten
  - Unicode
- Experimente
  - Hebräisch mit Nikkud
  - Schreibmaschinenschrift

Einleitung

Was ist synthetisches Training?

- Text-Bild-Alignierung als Grundlage jeden OCR-Trainings
  - Zeilen- oder Glyphenebene
  - hoher manueller Aufwand
  - hohe Fehleranfälligkeit

- Qualität der resultierenden Modelle abhängig von **Passung** und **Umfang** des Trainingsmaterials
- für die meisten Anwendungsszenarien: **keine** geeigneten Modelle und damit **schmutzige** OCR

Was ist synthetisches Training?

Training auf Basis von **maschinell generiertem** Trainingsmaterial

1. Font-Rendering-Engine
2. Font (digitale Fassung einer Schriftart)
3. Text

Text	Font	Bild
	Sans	B. A. Zimmermann
	Times	B. A. Zimmermann
B. A. Zimmermann	Claudius	B. A. Zimmermann
	Typenoksidi	B. A. Zimmermann
	Typenoksidi	B. A. Zimmermann

Wozu braucht man das?

- praktisch unbegrenzte Mengen an Trainingsmaterial generierbar
- Standardlösung für OCR-Training in `Tesseract 4`
  - Zitat: „400000 textlines spanning about 4500 fonts“
  - mglw. sogar einzige Möglichkeit des Trainings?
- Utilisierung manuell transkribierter Texte für OCR-Training (z. B. DTA)
- Kombination mit realen Trainingsdaten möglich
- bisher keine (systematische) Auseinandersetzung mit den einzelnen Parametern vor dem Hintergrund des OCR-Ergebnisses von gedrucktem Text

Technische Aspekte

Font Rendering

- Funktionalität von Betriebssystemen und Graphikbibliotheken
- unterschiedliche Techniken

  Bi-Level-Rendering: schwarze und „weiße“ Pixel

  Gray-Scale-Rendering: Modifikation der Helligkeit einzelner Pixel (*Anitaliasing*)

  Sub-Pixel-Rendering: Ansteuern der roten, grünen und blauen Subpixel → Verdreifachung der Auflösung

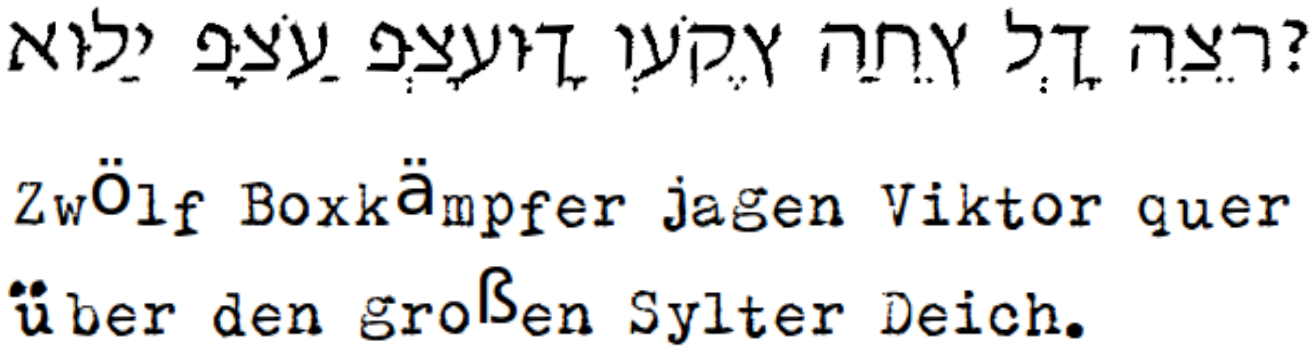
- viele verschiedene freie Implementierungen, z. B.
  - `ImageMagick`
  - `Python Image Library` (verwendet in `OCRopus`)
  - `Cairo` (verwendet in `Kraken`)

Schriftarten

- Implementierung einer Schrift
- unterschiedliche Techniken
  - `PostScript`
  - `TrueType`
  - `OpenType`
- für Font Rendering im Wesentlichen `OpenType` zu empfehlen
  - bessere **Unicode**-Unterstützung
  - bessere (automatische) Darstellung von **Ligaturen**
  - auch `*.ttf`-Schriften sind meist `OpenType`
- Ease-of-Use-Tools für die Erstellung **eigener Schriftarten**

Unicode

- Unicode-Unterstützung ≠ Unicode-Umsetzung
- gerade bei Font-Implementierungen
  - falsches Code-Point-Mapping (Windows)
  - unvollständige Zeichenumsetzung
- auch Issues beim Font-Rendering
  - Platzierung von Diakritika
  - automatischer Einsatz von Standardzeichen



Experimente

Experimente

Ziel: Ausloten der Potenziale des synthetischen Trainings in Kontexten ohne ausreichend vorhandenes reales Trainingsmaterial und mit erschwerten Bedingungen

Methode:

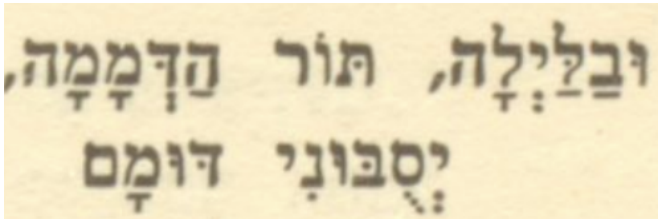
- Fontauswahl
- Generierung von Zeilenimages (unter verschiedenen Bedingungen) mit Hilfe von `OCRopus` und `Ketos`
  $\rightarrow$ nur binärcodierte Bilder
- Training von `OCRopus`- bzw. `CLSTM`-Modellen
- Evaluierung

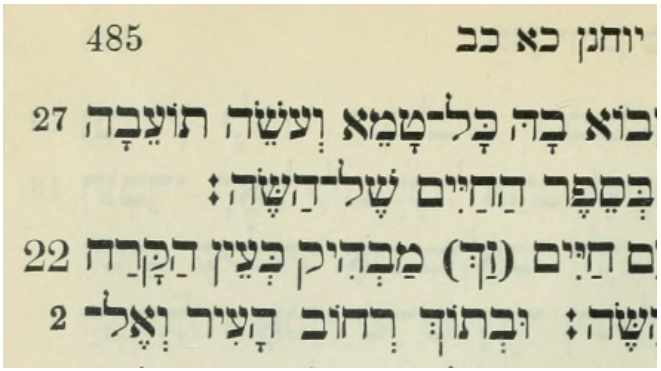
Experiment 1: Hebräisch mit Nikkud

- Hebräisch
  - alphabetische Sprache mit 22 Buchstaben
  - keine **Vokale**
- Nikkud („Punkte, Punktierung“)
  - verschiedene Diakritika (über, unter und **innerhalb** von Buchstaben)
  - Explifizierung von Vokalen und Konsonantenalternation
- Aufwand der manuellen Transkription **horrend**
- Texte insbesondere aus dem Bereich Lyrik und Drama aber auch Lexika
- Vorarbeiten mit Tesseract 3 (`https://www.cs.bgu.ac.il/~elhadad/hocr/`)

Experiment 1: Hebräisch mit Nikkud

Beispiele:


וּבַלַּיְלָה, תּוֹר הַדַּמָּה,
יְסַבּוּנִי דוֹמָם



27 בוא בָּהּ כָּל־טָמֵא וְעֹשֶׂה תּוֹעֵבָה
בְּסֵפֶר הַחַיִּים שֶׁל־הַשֶּׂה׃
22 ם חַיִּים (נֶ֫פֶשׁ) מַבְהִיק כְּעֵין הַקָּרְח
2 שֶׂה׃ וּבְתוֹךְ רְחוֹב הָעִיר וְאֶל־

מְהִיר־הָרַגְלַיִם נְחוּתוֹת אֲנִיוֹת הַמִּרְמִידוֹנִים. 20 בֵּין כָּל פְּתַלְתֹּל וְסָכָל לִבּוֹ כְּלֵב תַּחְכְּמוֹנִי.

Experiment 1: Hebräisch mit Nikkud

Vorgehen:

- manuelle Erfassung von zwei Seiten für Testzwecke (Text plus Layout, aligniert)
- Fontauswahl
  - wenige vollständig ausgestattete Schriften
  - **Frank-Rühl:** weltweit verbreitetste Schriftart für Hebräisch, 1909 von Rafael Frank in Leipzig entwickelt
  - ähnelt vielen Veröffentlichungen des 19. Jh.
  - darüberhinaus **KeterYG** und **DavidCLM**

Frank-Rühl	KeterYG	DavidCLM
בְּאֶמְצַע צַוָּארוֹ, וַיַּעֲבֹר וַיַּחֲלֹף חֹד־רָמְחוֹ הָאַדִּיר,	לְנַהֲגָן שְׁבִי בָאֲנִיָּה אֶל־אֶרֶץ מוֹלַדְתְּךָ אָהַבְתָּ.	מְטַנְּפִים בִּגְדֵיהֶם בִּדְמָעוֹת; וְהַשִּׁישׁ יוֹשֵׁב בַּתָּוֶךְ

Experiment 1: Hebräisch mit Nikkud

Vorgehen:

- Texttauswahl
  - unklarer Faktor
  - Saul Tschernichowskis hebräische Fassung der *Ilias*
- Kodierung der Buchstaben-Diakritika-Kombinationen als kombinierte Zeichen
  - Kodierung als Abfolge aus Buchstaben und Diakritika **untrainierbar**
  - 586 „Zeichen“
- Rendering mit `OCRopus` (verschiedene Grade der Degradation)
- Training mit `OCRopus` (Standardparameterbelegung)
  - nur synthetisches Trainingsmaterial
  - zwei Modelle: Frank Rühl only und Frank-David-Keter

Experiment 1: Hebräisch mit Nikkud

Ergebnisse:

- Transkription und Alignierung einer Seite Neues Testament sowie zwei Seiten Shakespeare
- jeweils 180000 Trainingsschritte, Auswahl des besten Modells

Modell	CER	Beschreibung
`Frank`	19 %	115000 Trainingsschritte
`FrankDavidKeter`	31 %	180000 Trainingsschritte
`Frank`	24 %	90000 Trainingsschritte
`FrankDavidKeter`	23 %	175000 Trainingsschritte

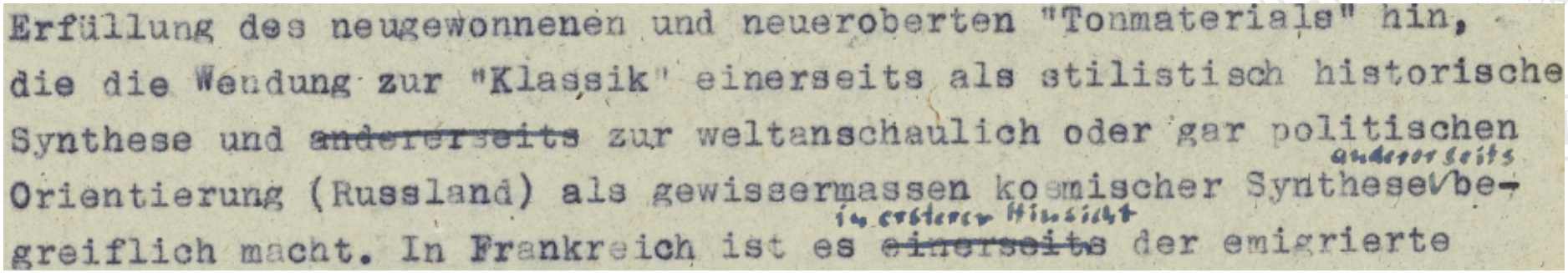
- Anwendung auf synthetisches Testmaterial: CER $< 2\,\%$

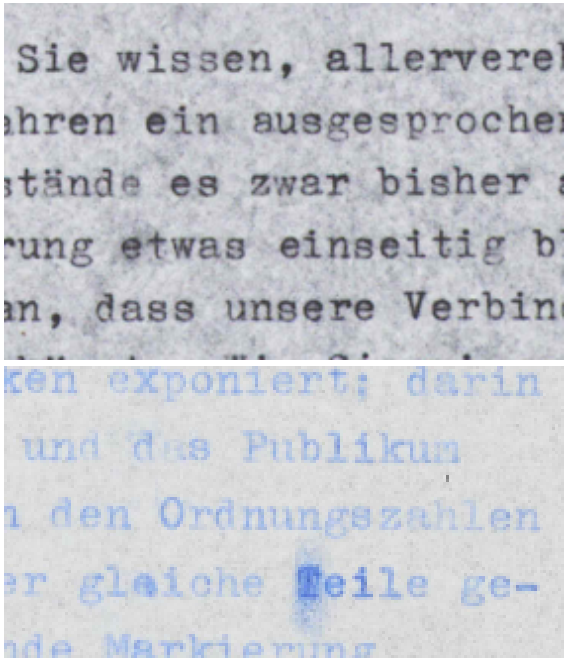
Experiment 2: Schreibmaschinentyposkripte

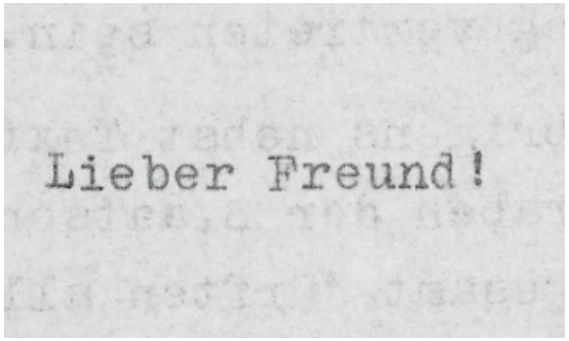
- im Kontext des Akademievorhabens Bernd Alois Zimmermann-Gesamtausgabe
- Volltexterfassung von ca. 6000 Seiten aus dem Nachlass, großer Anteil an **maschinengeschriebenen Briefen**
- Beschleunigung der Arbeit der Editoren durch initiale Textvorlage
- **schwierige Materialsituation** durch Durchschlagpapier, starke Materialalterung, handschriftliche Korrekturen
- Schreibmaschinentyposkripte prinzipiell gut für OCR geeignet
- wichtiges Anwendungsfeld (Archive!)

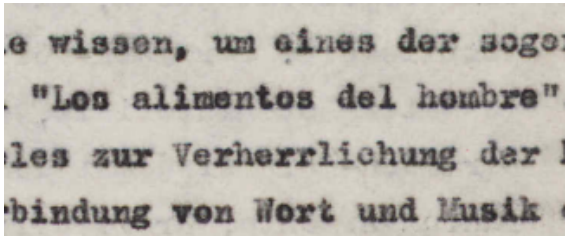
Experiment 2: Schreibmaschinentyposkripte

Beispiele:


Erfüllung des neugewonnenen und neueroberten "Tonmaterials" hin,
die die Wendung zur "Klassik" einerseits als stilistisch historische
Synthese und ~~andererseits~~ zur weltanschaulich oder gar politischen
Orientierung (Russland) als gewissermassen kosmischer Synthese ~~be-~~
greiflich macht. In Frankreich ist es ~~einerseits~~ der emigrierte


Sie wissen, allerverel
ahren ein ausgesproche
stände es zwar bisher
rung etwas einseitig b
an, dass unsere Verbin
ken exponiert; darin
und das Publikum
den Ordnungszahlen
er gleiche Teile ge-
nde Markierung


Lieber Freund!


e wissen, um eines der soge
"Los alimentos del hombre"
les zur Verherrlichung der
bindung von Wort und Musik

Experiment 2: Schreibmaschinentyposkripte

Vorgehen:

- manuelle Erfassung von zehn Seiten für Testzwecke (Text plus Layout, aligniert)
- Fontauswahl
  - riesige Menge an vorhandenen Schriften
  - teilweise modellspezifisch
  - mit entsprechenden Artefakten

B. A. Zimmermann	B. A. Zimmermann	**B. A. Zimmermann**	B. A. Zimmermann
1913 Underwood No. 5	Rheinmetall 1952	Byron Mark 1	Olivetti Lettera 32
			

Experiment 2: Schreibmaschinentyposkripte

Vorgehen:

- Texttauswahl
  - unklarer Faktor
  - intuitive Auswahl: Briefe
  - Jean Paul: „Dritte Abteilung Briefe“. In: *Jean Pauls Sämtliche Werke. Historisch-kritische Ausgabe*. Abt. 3, Bd. 1. Berlin, 1956.
- Rendering mit `Ketos` (`--legacy`-Degradation)
- Training mit `CLSTM` (Standardparameterbelegung)
  - nur reales Trainingsmaterial
  - nur synthetisches Trainingsmaterial
  - gemischt

Experiment 2: Schreibmaschinentyposkripte

Ergebnisse:

- 215 Zeilen reales Trainingsmaterial, 95 Zeilen Testmaterial
- jeweils 40000 Trainingsschritte, Auswahl des besten Modells

Modell	CER	Beschreibung
`Abbyy`	5 %	Abbyy FineReader 14
`Real`	17 %	17000 Trainingsschritte
`Synthetic`	25 %	21000 Trainingsschritte auf ca. 9000 Zeilen
`Mixed-simple`	22 %	17000 Trainingsschritte auf ca. 10000 Zeilen
`Mixed-komplex`	15 %	26000 Trainingsschritte auf ca. 9000 Zeilen

- Anwendung auf synthetisches Testmaterial: CER $< 1\,\%$

Hypothesen und Lösungsmöglichkeiten

Hypothesen und Lösungsmöglichkeiten

- Distanz zwischen synthetischem Trainingsmaterial und realen Daten zu groß

  → mehr bzw. realistischere Degradation

  → passendere Schriftarten

  → realistischeres Rendering (Graustufen oder Farbrendering)

  → bessere Vorverarbeitung

- Modelle konvergieren zu schnell

  → mehr Trainingsdaten, mehr Schriftarten

  → alternative Parameterbelegung beim Training

- Hebräisch: Alphabet zu groß

  → getrennte Erkennung von Buchstaben und Diakritika

Danke für Ihre Aufmerksamkeit!

OCR-D-Team: Elisa Hermann, Maria Federbusch, Clemens Neudecker,
Ajinkya Prabhune, **Matthias Boenig**