Erstellen von Trainingsdaten mit Franken+

1. Was ist Franken+?

Franken+ ist ein *Tesseract*-Frontend, dass an der Texas A&M University entwickelt wurde. Das Tool ist darauf spezialisiert, das Erstellen von Tesseract-Trainingsdaten zum OCR von historischen Schriften zu erleichtern. Hierzu wird der Output von PRImALab's *Aletheia* verwendet, welcher im Rahmen dieser Präsentation schon erstellt wurde. Franken+ verarbeitet das PAGE XML in eine Schriftart mit einzelnen Glyphen, welcher wiederum zur Erstellung von synthetischen Bilddateien eingesetzt wird, die schließlich die Grundlage für das Training mit Tesseract darstellen. In diesem Vortrag stellen wir den aktuellen Stand der Forschung in der stilometrischen Autorschaftsattribution mit Delta und seinen Varianten vor und berichten neue Beobachtungen und Erkenntnisse aus eigenen Untersuchungen.

2. Workflow-Übung Franken+

Da Franken+ seine Datenverwaltung mit einer MySQL-Datenbank managed, muss diese bei der ersten Nutzung des Programms erst verknüpft werden. Sollte es Grund geben die Datenbank im Nachhinein zu verändern, kann diese unter dem Reiter *Menu* → *Settings* → *Database* getan werden.

Da Franken+ keine eigene Tesseract-Version mitliefert, muss ebenfalls vor der Nutzung angegeben werden, in welchem Pfad die Tesseract-Installation zu finden ist. Nachdem das Programm korrekt konfiguriert wurde, ist der erste Schritt einen neuen Font zu erstellen, was durch den *Create*-Button initialisiert werden kann. Im nun geöffneten Fenster ist zu Beginn lediglich der Font-Name einzugeben und eventuell die Font-Variante (Italic, Bold, …) auszuwählen, wobei letztere keinen Einfluss auf das OCR-Ergebnis hat, sondern nur Folgen für die Visualisierung hat. Da die anderen Optionen ohne Glyphen noch keine Bedeutung besitzen, kann der Font nun via *Save Font* gespeichert werden. Ist der gewünschte Font nun im Drop-Down-Menü ausgewählt, kann unter *Aletheia TIF/XML* der Ordner ausgewählt werden, in dem die mit Aletheia erstellten TIF/XML-Paare zu finden sind. Hierbei muss beachtet werden, dass die TIF und die XML Dateien – bis auf die Endungen – identische Namen besitzen.

Wurde bei der Erstellung ein oder mehrere Glyphen durchgehend dem falschen Unicode zugeordnet, so kann dies durch die *Unicode Substitutions* umgangen werden. Hierzu wird ein Häkchen bei *Use with ingestion* gesetzt und unter *Create New* die entsprechende Ersetzungsregel eingegeben. Mit dem Button *Ingest Glyphs* werden diese nun in einzelne Glyphen aufgeteilt und im erstellten Font den passenden Unicode-Characters zugeordnet. Dieser Vorgang kann je nach Masse an vorbereiteten TIF/XML-Paaren und der Rechenleistung einige Zeit in Anspruch nehmen. Wurde das *Ingesten* erfolgreich abgeschlossen, kann das Ergebnis im Font betrachtet werden. Öffnet man den Font nun

mit *Edit*, so sind im Dropdown-Menü unter Glyphs die passenden Glyphen als einzelne Bilder dem passenden Unicode-Zeichen zugeordnet.

Ein Hauptteil der Arbeit bei der Erstellung von Trainingsdateien fällt nun darauf, dass fehlerhafte Glyphen aussortiert werden und die Glyphen gefunden werden, welche das beste Ergebnis liefern. Um alle möglichen Probleme bei der Arbeit mit den Glyphen abzudecken, wird im Rahmen dieses Workshop zunächst eine vorbereiteter Font importieren, in welchem typische Fehler eingebaut wurden. Hierzu wird unter *Menu → Import Font* die entsprechende ZIP-Datei ausgewählt und dem Font ein beliebiger Name zugewiesen. Der Export eines Fonts verläuft ähnlich unkompliziert unter *Menu → Export Font.* Nun kann mit der Arbeit an den Glyphen und dem Font begonnen werden.

Grundsätzlich ist hierbei die Maxime, Glyphen von Schmutz zu bereinigen und deutlich fehlerhafte Glyphen zu entfernen. Hierbei ist jedoch anzumerken, dass keine Anleitung für das Vorgehen zum Erstellen eines perfekten Fonts gegeben werden kann. Vor allem die letzten Prozent bei der Erkennungsrate wurden oft durch schlichtes *Trial and Error* herausgekitzelt. So war es teilweise zielführend sich an die offiziell als ideal bezeichnete Menge von fünf bis maximal zehn Glyphen zu halten, in anderen Fällen führten eins bis zwei oder mehr als zehn Glyphen zu besseren Ergebnissen. Auch das Behalten von Glyphen, welche von der Norm abwichen, brachte teilweise Verbesserungen, verschlechterte jedoch auch in vielen Fällen das Ergebnis.

Soll eine Glyphe aus dem Font ausgeschlossen werden, so besteht zum einen die Möglichkeit diesen lediglich als fehlerhaft zu markieren, was durch ein Klicken auf den Glyphen erreicht wird. Ist die Glyphe rot, so ist sie nicht aus der Datenbank gelöscht, wird jedoch nicht mehr beim Training berücksichtigt. Dies ist vor allem bei Tests sinnvoll. Sollen Glyphen jedoch endgültig gelöscht werden, so werden Glyphen rot markiert und mit dem Button *Delete Removed* gelöscht. Ein erneuter Klick auf *Removed Glyhps* lässt diese wieder grün erscheinen wodurch sie wieder als normal gelten. Zu beachten ist, dass Franken+ die Glyphen auf mehrere Seiten aufteilt, welche durch *Prev Page* und *Next Page* angesteuert werden können. Mit einem Rechtsklick auf eine Glyphe und *Edit Image* besteht die Möglichkeit die Bilddatei mit einem beliebigen Bildbearbeitungsprogramm (zum Beispiel Photoshop) zu bearbeiten und abzuspeichern. Dies ist vor allem bei der Entfernung von Schmutz hilfreich. Mit diesem Wissen kann nun begonnen werden, den Font in der Praxis auf das spätere Training vorzubereiten.

Beispiele fehlerhafter Glyphen

- S: Defekte Glyphe → Reparieren mit Photoshop
- e: Unsauberer Druck/Scan/Digitalisierung
- h: Keine Offsetanpassung
- i: Nur Teile einer Glyphen → Entsteht durch fehlerhafte Polygonisierung in Aletheia
- p: Stark unterschiedliche Größe der Glyphen → Kann zu Fehlern führen, muss jedoch nicht!
- s: Typische Häufung verschiedener Fehler nach dem Ingesten
- t: Abwägung welche Glyphen dem kleinen r zu ähnlich sehen
- ħ: Zwei Unicode Characters → Probleme beim Erzeugen synthetischer TIF/BOX-Paare

Ist der Font nun fertig bearbeitet, ist der nächste Schritt die synthetischen *TIF/BOX*-Paare zu erzeugen, welche später die Grundlage für das Training darstellen. Hierzu wird zuerst eine *Language* erstellt, was durch den *Create Language* Button erreicht wird. Ebenfalls notwendig ist mindestens eine Textdatei, welche als Grundlage für das zu erzeugende Bild gilt. Von Franken+ selbst werden Transkriptionsdateien als *Ground Truth* empfohlen. Das beste Ergebnis wurde in unseren Anwendungsfällen jedoch beim Zusammenspiel von Transkriptionen mit einer Textdatei erzielt, welche jeden im Font vorkommenden Glyphen mindestens einmal enthielt und diese im Text zufällig anordnete. Dies ist vor allem bei Schriftarten sinnvoll, die viele verschiedene Sonderzeichen enthalten, welche jedoch nur vereinzelt im Text vorkommen. Durch die zufällig angeordnete Textdatei wird das mögliche Problem abgefangen, dass diese Sonderzeichen nicht im Transkriptionstext vorkommen, wodurch diese nicht beim späteren Training berücksichtigt werden würden. Bevor nun mit dem Training begonnen werden kann, müssen die erzeugten synthetischen TIF/BOX-Paare überprüft werden und eventuell Anpassungen am Font gemacht werden. Neben dem Erkennen von defekten Glyphen welche zuvor übersehen wurden, ist vor allem auf die Baseline der Glyphen zu achten.

Vor allem bei Buchstaben, die teilweise unterhalb der Baseline sind *(g, j, …)* muss in den meisten Fällen der vertikale Offset angepasst werden. Dies geschieht wieder im Font-Fenster und ist nach der Auswahl eines bestimmten Unicodes unter *Selected Glyph Offsets* einstellbar. Die eingegebenen Zahlenwerte verschieben die Glyphen pixelweise. Vor allem der Y-Offset ist hier von Interesse. Für diese Einstellungen bestehen keine genauen Richtlinien, deshalb muss der perfekte Offset durch Probieren und Kontrollieren der synthetisch erzeugten TIF/BOX-Paare herausgefunden werden.

Sollten diese nun zur Zufriedenheit angepasst worden sein, kann schließlich mit dem Training begonnen werden. Hierzu wählt man die gewünschte Sprache aus dem Dropdown-Menü aus und klickt auf *Train Tesseract.* Im sich nun öffnenden Fenster werden verschiedene Einstellungsmöglichkeiten angeboten. Auf der linken Seite werden die Fonts ausgewählt, welche zum Training hinzugezogen

werden sollen. Das Training von bis zu drei verschiedenen Fonts führte zu annehmbaren Ergebnissen, bei einer höheren Anzahl kam es jedoch zu erheblichen Problemen bei der späteren Texterkennung. Empfohlen wird das Training eines einzelnen Fonts und die vorherigen Segmentierung des Werkes auf das die OCR angewendet werden soll. Hierdurch wurde das beste Ergebnis erzielt. Unter *files to include* besteht die Möglichkeit *unicharambigs* und *Dictionaries/Word Lists* in die Trainingsdatei einzufügen. *Unicharambigs* ermöglichen zum einen das erzwungene Ersetzen von bestimmten Unicodes und Unicodeabfolgen, was zum Beispiel bei bestimmten Ligaturen Sinn macht, andererseits aber auch das Abwägen der Ersetzung anhand eines Wahrscheinlichkeitswertes und eines Wörterbuches. Der Modus der *Unicharambigs* bietet zwar unglaublich viele Möglichkeiten, konnte jedoch bei unseren Arbeiten nicht sinnvoll eingesetzt werden, da Probleme bei der Benutzung bestehen.

Die Word Lists können Franken+ als einfache .txt-Dateien zur Verfügung gestellt werden. Vor allem die *Frequent Word List* besaß teilweise starke Auswirkung auf die Erkennungsrate. Problematisch ist jedoch vor allem bei der Bearbeitung von historischen Texten, dass selten ein passendes Wörterbuch vorhanden ist. Dies wird durch unterschiedliche Schreibweisen und die diversen Sonderzeichen nahezu unmöglich gemacht. Unter dem Reiter *Training Steps* ist in den meisten Fällen lediglich die Checkbox *Clear old training* zu beachten, welche bei mehrmaligen Training derselben Language demarkiert werden sollte. Mehrmaliges Training führte in vielen Fällen zu einer Verbesserung der Ergebnisse, eine genau eAnzahl an idealen Trainingsdurchläufen kann jedoch auch hier nicht angegeben werden. Sind alle Einstellungen getroffen, wird das Training durch *Make Library* gestartet. Dies kann je nach Umfang der TIF/BOX-Paare von weniger als einer Minute bis zu mehreren Stunden dauern. Ist das Training beendet, öffnet sich der Ordner mit den Trainingsdateien, bei welcher vor allem die *.traineddata* für OCR mit Tesseract wichtig ist. Diese muss zur Verwendung in den *traineddata*-Ordner von Tesseract abgelegt werden.

Maximilian Nöth

JMU Würzburg