Know your Data: KI-Anwendungen passgenau konfigurieren

Annette Gerstenberg[1], Hanno Müller[2], Friederike Schulz[1]

Für die Verschriftung von Tonaufnahmen werden Anwendungen der Künstlichen Intelligenz (KI) genutzt, nicht nur in der Wissenschaft. Das spart Zeit, aber bei spontaner Sprache und stimmlichen Besonderheiten ist die zeitaufwändige manuelle Korrektur automatisch erstellter Transkripte notwendig. Unser Beitrag stellt mit Beispielen aus dem Projekt LangAge[1] und Evaluationsergebnissen einen Workflow zur verbesserten Transkription französischer Interviews mit älteren Sprecherinnen und Sprechern vor. Anschließend wird mit dem Publikum diskutiert, wie der vorgeschlagene Weg der Konfiguration von KI-Anwendungen und der Datenvorbereitung auf andere Datentypen übertragen werden kann.

1 Anpassung 1: Das Transformer-Modell FrWhisper

Problem1: Im Projekt LangAge wurde zur Kontrolle manueller Transkripte Whisper large-v3 verwendet, mit unbefriedigendem Ergebnis bzgl. der Merkmale der gesprochenen Sprache (Wiederholungen, Füllwörter) und alterstypischer Merkmale, bekannte Bias-Phänomene.

Lösung1: In Kooperation mit dem Hasso-Plattner-Institut Potsdam wurde mit Korpus-Daten durch Fine-Tuning von Whisper large-v3 ein neues Transformer-Modell erstellt: FrWhisper. Es kann lokal genutzt werden und ist für Eigenheiten speziell des gesprochenen Französischen und speziell für Stimmen älterer Sprecher trainiert. Der Fortschritt zeigt sich in der Evaluation: Im Vergleich zu Whisper large-v3 lag die Fehlerquote (WER) in FrWhisper-Transkripten für die Korpusdaten um gut 20 Prozentpunkte niedriger.[3]

2 Anpassung 2: Qualitätssprung durch Datenvorbereitung

Problem2: In FrWhisper-Transkripten werden Wiederholungen, die tatsächlich vorkommen, in überlangen Schleifen wiederholt. Diese Schleifen treten vor allem auf, wenn FrWhisper auf stundenlangen Tondateien läuft.

Lösung2: Die Tondateien werden in passgenaue Einzeldateien geschnitten. Ein lokal laufendes neuronales Netz zur Sprachaktivitätserkennung (Voice Activity Detection, VAD) erzeugt eine Datei (*.TextGrid), in der sprachliche von nicht-sprachlichen Passagen getrennt werden. Damit wird die Tondatei geschnitten. Die Qualität der FrWhisper-Transkripte der geschnittenen Dateien steigt, Wiederholungsschleifen treten nur noch vereinzelt auf – das erbringt eine deutliche Zeitersparnis bei der manuellen Überarbeitung.

Ablauf: 15 Minuten Präsentation, 15 Minuten Diskussion.

---

[1] Projekt LangAge, Lehrstuhl Sprachwissenschaft Französisch und Italienisch, Universität Potsdam.

[2] Hasso-Plattner-Institut Potsdam, Projektförderung BMFTR 01IS22092.

[3] Modellkarte und Beschreibung des Modells auf Huggingface: https://huggingface.co/aihpi/FrWhisper.