Oachkatzl

Training und Benchmarking von KI-basierten Audio-Mining-Systemen auf bayerische Dialekte

Johannes Lederle

Abstract

Es gibt inzwischen fast niemanden mehr, der den Begriff „künstliche Intelligenz“ noch nicht gehört hat. Doch wie kann diese Technologie für den eigenen Anwendungsfall genutzt werden? Das Projekt „Oachkatzl“ verfolgt das Ziel, die automatische Transkription von bayerischen Dialekten zu verbessern, um die Einsatzmöglichkeiten von KI-basierten Audio-Mining-Tools beim BR zu erweitern. Am Beispiel dieses Projekts wird der Prozess vorgestellt, der ein Audio-Mining-Tool auf diesen speziellen Anwendungsfall trainiert. Größte Herausforderung ist dabei der Aufbau einer möglichst hochwertigen Trainingsdatenbank. BR-eigene Daten und bereits existierende Metadaten (z.B. Untertitel) können hier ein großes Wertschöpfungspotential bieten. Darüber hinaus wird die Frage beantwortet, ob ein teilautomatisierter Workflow zur Trainingsdatengenerierung bei der Entwicklung eines Audio-Mining-Tools von Nutzen sein kann.

Einführung

Die rapide voranschreitende Digitalisierung und Globalisierung führen zu einem exponentiellen Anstieg der Produktion, Speicherung und Veröffentlichung von Daten aller Art in allen Bereichen unseres Lebens. 2019 wurden weltweit über 30 Milliarden Terrabyte an Daten produziert, bearbeitet oder veröffentlicht[1]. Ein großer Anteil dieser Daten ist besonders für die Medienbranche interessant (u. a. Presse/Onlinemeldungen und Audio- oder Videobeiträge). Entscheidend für einen reibungslosen und effektiven Arbeitsablauf in einem Medienunternehmen ist die schnelle Analyse des Informationsstroms und die Verarbeitung der richtigen Daten zu relevanten Beiträgen. Dabei spielt die Erschließung von Informationen aus natürlicher Sprache (u.a. aus Audio- und Videodateien) eine zentrale Rolle. Wegen der Zunahme der Informationsmenge ist eine automatische Verarbeitung der Daten unumgänglich. Nur so kann eine ausreichend breite und detaillierte inhaltliche bzw. formale Erschließung der Daten mit akzeptablem Ressourcenaufwand stattfinden. Hier bieten künstliche Intelligenz basierte (KI) Audio-Mining-Systeme (AM) großes Potenzial. Diese Tools überführen Sprache in geschriebenen Text. Anschließend kann über weitere Verfahren der Inhalt weiter aufbereitet werden.

Johannes Lederle
Technische Universität München/Bayerischer Rundfunk
Albert-Einstein-Straße 12,
85375 Neufahrn
johannesslederle@gmail.com

Ein produktiv einsetzbares KI-System setzt jedoch eine große Menge an qualitativ hochwertigen Trainingsdaten voraus. Ein Trainingsdatensatz für die AM-Anwendung besteht aus einem kurzen Audiofile mit natürlicher Sprache (Chunk) und einem Textdokument mit den Wörtern, die gesprochen wurden (Transkript). Für die Entwicklung eines produktiv einsetzbaren AM-Tools werden etwa 1000 Stunden Trainingsmaterial (bestehend aus vielen einzelnen Trainingsdatensätzen) benötigt. Beim Aufbau solch einer Trainingsdatenbank steht man insbesondere vor zwei Herausforderungen. Zum einen müssen die Audiofiles mit geeigneter Sprache vorhanden sein oder extra erstellt werden. Zum anderen müssen die Transkripte händisch erstellt werden. Diese Aufgabe ist äußerst zeit- und kostenintensiv.

Aktuell sind qualitativ hochwertige AM-Systeme für hochdeutsche Sprache erhältlich. Für bayerische Dialekte existiert aktuell kein Produkt, welches akzeptable Ergebnisse liefert.

Der Bayerische Rundfunk (BR) hat als regionale Rundfunkanstalt einen großen Anteil an bayerischer Sprache im Medienangebot. Bei der automatischen Analyse bayerischer Dialekte mit aktuell verfügbaren Systemen enthalten die Transkripte jedoch viele Fehler. Dies schränkt den breiten Einsatz von AM-Tools zum jetzigen Zeitpunkt stark ein.

In den Datenbanken des BRs befinden sich große Mengen an Fernsehsendungen und Beiträgen mit

1 R. T. Kreutzer und M. Sirrenberg, Künstliche Intelligenz verstehen. Wiesbaden: Springer Fachmedien Wiesbaden, 2019

bayerischen Sprechern. Darüber hinaus existieren zu einem Teil dieser Videos Untertitel. Diese könnten als Grundlage für eine Trainingsdatenbank dienen.

Im Rahmen des Proof of Concept „Oachkatzl“ wurde geprüft, ob die Sendungen in Chunks und die Untertitel in Transkripte verarbeitet werden können, um ein AM-System erfolgreich auf bayerischen Dialekt zu trainieren.

Ziel ist es, die Erfolgsaussichten und den Arbeitsaufwand für die Entwicklung eines bayerischen AM-Systems mit BR-eigenen Archivdaten abschätzen zu können.

Das Projekt „Oachkatzl“ wurde beim BR von der Abteilung „Dokumentation und Recherche“ in enger Abstimmung mit der Produktions- und Technikdirektion und der Redaktion „Barrierefreie Angebote“ initiiert. Unterstützt wurde das Projekt durch das Softwareunternehmen „ainblick“, geführt von Olaf Thiele. „ainblick“ lieferte Software-Tools und Arbeitskraft zur Vorverarbeitung der Trainingsdaten und übernahm das Training der AM-Systeme.

Neuronale Netze – Das Herzstück der KI

Die Aufgabe, natürliche Sprache zu verstehen und zu verschriftlichen, ist eine große Herausforderung, da jede Person andere Sprechgewohnheiten (z.B. durch Dialekte) hat, gleiche Wörter unterschiedliche Bedeutungen haben können oder Störgeräusche die Analyse erschweren. Deshalb werden bei der Spracherkennung Konzepte aus dem Bereich der künstlichen Intelligenz bzw. dem maschinellen Lernen angewendet. Aktuelle AM-Systeme arbeiten in den meisten Fällen mit künstlichen neuronalen Netzen.

Künstliche neuronale Netze sind den neuronalen Netzen des menschlichen Gehirns nachempfunden. Ein neuronales Netz besteht aus vielen einzelnen Neuronen, die untereinander unterschiedlich stark vernetzt sind. Bei der Signalverarbeitung werden die Neuronen der Input-Schicht (vgl. Abb. 1 „h1“) angeregt. Diese geben je nach Signalstärke einen Impuls an die nächsten Neuronen weiter, welche wiederum Impulse an die folgenden Schichten abgeben, bis die letzte Schicht das gewünschte Analyseergebnis ausgibt. Für die korrekte Informationsverarbeitung ist entscheidend, welche Neuronen angeregt werden und welche nicht. Dies lässt sich unteranderem über die Parameter „Gewichtung“ und „Bias“ beeinflussen, welche durch das Training angepasst werden.

Für das hier beschriebene Projekt wird die Open-Source-Software Deep Speech des Unternehmens Mozilla eingesetzt. Deep Speech basiert auf einem rekurrenten neuronalen Netz, welches sich mit eigenen Trainingsdaten trainieren lässt (vgl. Abb. 1).

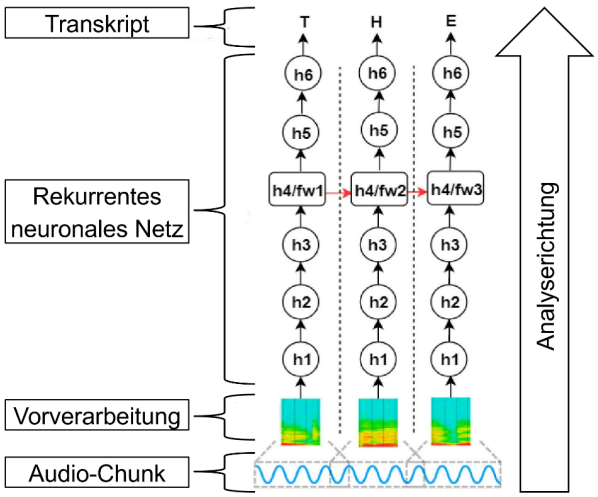

Abbildung 1: Schematische Darstellung des Deep Speech Modells (in Anlehnung an: T. Z. Aashish Agarwal, „German End-to-end Speech Recognition based on Deep-Speech“, 2019. Verfügbar unter: https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_23.pdf)

Training von AM-Systemen

Zentrale Aufgabe bei der Entwicklung eines produktiv einsetzbaren AM-Tools ist die Beschaffung von hochqualitativen Trainingsdaten. Wie zu Beginn erwähnt, besteht eine geeignete Trainingsdatenbank aus einer großen Anzahl an Chunks mit dazu passenden Transkripten (vgl. Abb. 2).

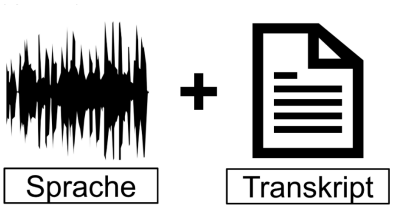

Abbildung 2: Trainingsdatensatz bestehend aus Chunk und Transkript

Erfahrungen der Vergangenheit haben gezeigt, dass ein AM-System ab einem Training mit etwa 500-1000 Stunden Material für den praktischen Einsatz ausreichend gute Ergebnisse liefert (korrekte Trans-

kription > 80 %). Falls nicht genügend Trainingsdaten verfügbar sind, kann ein Nachtraining bessere Ergebnisse liefern als ein vollständiges Training. Bei einem Nachtraining wird ein bereits auf einen sehr ähnlichen Anwendungsfall vollständig trainiertes neuronales Netz mit einem kleinerem anwendungsspezifischen Trainingsdatensatz nachtrainiert (z. B. Nachtraining eines hochdeutschen AM-Tools auf bayerische Dialekte).

Die Beschaffung von geeigneten Audio-Files mit passender gesprochener Sprache und die Erstellung der korrekten Referenz-Transkripte ist eine Aufgabe, die einen großen Arbeitsaufwand bedeutet. So können laut „ainblick“ für ein KI-Projekt in der Regel 80% der Ressourcen für die Beschaffung und Erstellung der Trainingsdaten und 20 % für das eigentliche Training und Tests einkalkuliert werden.

Aus diesem Grund ist eine möglichst effiziente Trainingsdatengenerierung, zum Beispiel durch teilautomatisierte Arbeitsschritte und die Verwendung spezieller Annotations-Software entscheidend für einen akzeptablen Ressourcenaufwand beim Aufbau einer Trainingsdatenbank.

Word-Error-Rate – Kennzahl der Transkriptionsqualität

Für den Benchmark verschiedener AM-Systeme müssen die Output-Qualitäten (d.h. die Transkripte) verschiedener Modelle präzise miteinander verglichen werden. Dafür bietet sich die Wortfehlrate (engl. Word-Error-Rate (WER)) sehr gut an. Für die Bestimmung der WER benötig man zusätzlich zum Transkript des AM-Systems ein Referenz-Transkript (bzw. eine Musterlösung), welches die geforderte Ausgabe eines AM-Tools angibt. Die WER gibt an, wie viel Prozent der Wörter des AM-Transkripts vom Referenz-Transkript abweichen. Die Formel zur Berechnung der WER lautet wie folgt:

$$WER = \frac{\#Korrektur + \#Entfernung + \#Ergänzung}{\#\,Wörter\ in\ der\ Referenz}$$

$\#Ersetzung = Anzahl\ Wörter, die\ korrigiert\ werden\ müssen$
$\#Entfernung = Anzahl\ Wörter, die\ zu\ viel\ verschriftlicht\ wurden$
$\#Ergänzug = Anzahl\ Wörter, die\ nicht\ verschriftlicht\ wurden$
$\#Wörter\ in\ der\ Referenz = Anzahl\ Wörter\ des\ Referenz - Transkripts$

Für einen Benchmark eines AM-Systems gilt, je niedriger die WER desto weniger Fehler sind im automatisch erzeugten Transkript und desto besser ist ein System produktiv einsetzbar.

Projekt Oachkatzl

Für das hier durchgeführte Projekt sollen die Trainingsdaten aus bayerisch-dialektal gefärbter Sprache bestehen. Im Detail soll dialektale Sprache in hochdeutschen Text transkribiert werden. Aus diesem Grund müssen die Trainingsdaten in dieser Form vorliegen (vgl. Abb. 3).

Gesprochener Satz:
„Oder i ham zugehert wie mei Opa mit der Schüler etwas erklärt hat.“

Ideales Trainings-Transkript:
„Oder ich habe zu gehört wie mein Opa mit der Schüler etwas erklärt hat.“

Abbildung 3: Beispiel eines idealen Trainingsdatensatzes

In Abbildung 3 ist besonders auffällig, dass grammatikalische Fehler übernommen werden (vgl. „mit der Schüler“). Ziel dieses Projekts ist zunächst eine korrekte „Wort für Wort“-Transkription. Fehler im Satzbau und in der Grammatik können nachträglich durch Tools, ähnlich der Word-Rechtschreibprüfung, verbessert werden.

Die Produktion der oben beschriebenen Datensätze ist äußerst aufwändig. Eine vollständige Transkription benötigt etwa die dreifache Zeit der Abspieldauer der Daten. Das bedeutet für 500 Stunden Trainingsmaterial müssten 1.500 Stunden Arbeitszeit kalkuliert werden. Um den Zeit- und Kostenaufwand für die Erstellung der Trainingsdatenbank zu senken, werden Daten des BRs verwendet, die einem Trainingsdatensatz bereits ähneln. Diese müssen lediglich überarbeitet werden, was weniger Aufwand bedeutet.

Hierbei bietet das BR-Archiv enormes Potenzial (vgl. Abb. 4).

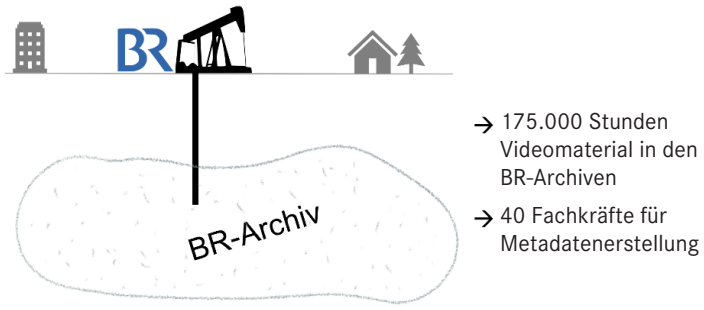

Abbildung 4: Das BR-Archiv – Eine „Ölquelle“ für Trainingsdaten

Besonders zwei Faktoren machen das BR-Archiv wertvoll für die Trainingsdatengenerierung. Zum einen befinden sich allein 175.000 Stunden Videoma-

terial in den Archiven. Ein Teil dieses Materials kann als Chunks für das AM-Training verwendet werden. So entfällt der Aufwand, Audio-Material selbst aufzunehmen. Zum anderen haben etwa 40 Fachkräfte beim BR die Aufgabe, neu produziertes Material mit Metadaten zu versehen. Dies ist erforderlich, um die Recherchierbarkeit in den Archiven zu gewährleisten. Für das AM sind Untertitel besonders relevant, da diese die gesprochenen Inhalte in Schriftform wiedergeben. Fernsehsendungen mit Untertiteln können so als Grundlage für die Trainingsdatensätze dienen.

Aufbereitung der Trainingsdaten

Um Fernsehsendungen mit Untertiteln möglichst effektiv zu qualitativ hochwertigen Trainingsdaten zu verarbeiten, wurde in Zusammenarbeit von BR und „ainblick“ ein teilautomatisierter Prozess entworfen (vgl. Abb. 5).

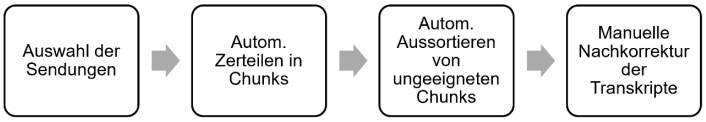

Abbildung 5: Prozess zur Generierung von Trainingsdaten aus Fernsehsendungen und Untertiteln

Zu Beginn werden Fernsehsendungen aus den BR-Archiven ausgewählt, die am besten für ein AM-Training geeignet sind. Die optimalen Anforderungen an die Sendungen lauten:

- Sendungslänge mindestens 20 Minuten lang, um möglichst viel Rohmaterial in möglichst geringer Anzahl an Dateien zur Verfügung zu haben
- Möglichst viele verschiedene bayerische Sprecher*innen mit unterschiedlichen bayerischen Dialekten
- Sendung muss Untertitel besitzen

Im Süden von Deutschland und in Teilen Österreichs wird eine große Anzahl an verschiedenen Dialekten gesprochen, welche dem „Bayerischen“ zugeordnet werden können. Jedoch ist die klare Abgrenzung einzelner Dialekte in den meisten Fällen nicht möglich. Aus diesem Grund werden für dieses Projekt alle bayerischen Dialekte unsortiert in die Trainingsdatenbank aufgenommen (vgl. Abb. 6). Dies senkt zum einen den Aufwand des Dialekt-Sortierens, zum anderen sollte das AM-System nach erfolgreichem Training, jeden bayerischen Dialekt korrekt transkribieren können.

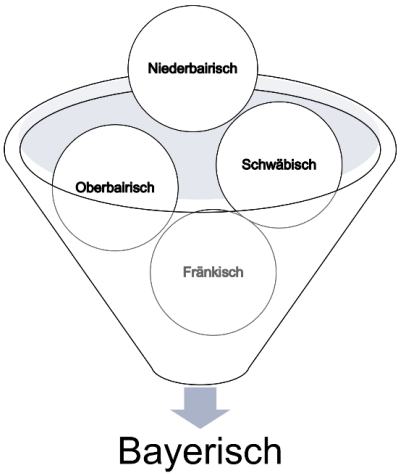

Abbildung 6: Zusammenfassung aller bayerischen Dialekte zu einem Hauptdialekt für das Training

Besonders geeignet für die Roh-Trainingsdaten, sind die Sendungen „Dahoam is Dahoam“, „Komödienstadl“, und „Unter unserem Himmel“. Insgesamt wurden etwa 237 Stunden Videomaterial mit Untertiteln für dieses Projekt gesammelt.

Im zweiten Schritt der Trainingsdatengenerierung werden die ausgewählten Sendungen in Chunks zerteilt. Dafür stellt die Firma „ainblick“ ein Tool bereit, welches teilautomatisiert das Videofile in 6-8 Sekunden Abschnitte unterteilt, die Untertiteldatei zerteilt und dem richtigen Chunk zuordnet.

Anschließend müssen nicht geeignete Chunks aussortiert werden. Chunks ohne Sprache, mit Hintergrundgeräuschen oder mit 2 Sprechern, die gleichzeitig reden können, nicht für das Training verwendet werden. Da besonders Serien wie „Dahoam is Dahoam“ immer dem gleichen Ablaufplan folgen (Rückblende mit Hintergrundmusik, Vorspann, Handlung, Abspann, etc.), kann hierfür ein Tool verwendet werden, welches Chunks aus ungeeigneten Passagen automatisch entfernt. Dies senkt den manuellen Sortieraufwand immens.

Größten manuellen Aufwand benötigt die Nachkorrektur der Transkripte, da sich Untertitel trotz ähnlicher Funktion von Transkripten unterscheiden. Grundsätzlich kann ein Mensch schneller gesprochene Wörter wahrnehmen, als er diese lesen kann. Damit durch Untertitel in der gleichen Zeit dieselbe Menge an Informationen wiedergegeben werden können, wie durch gesprochene Sprache, müssen die Untertitel oft sprachlich und inhaltlich gerafft werden. Dies hat zur Folge, dass die Untertitel keine präzise Transkription (Wort für Wort) der

Sprache bieten. Da exakte Transkripte enorm wichtig für hochqualitative Trainingsdaten sind, müssen die Untertitel nachbearbeitet werden (vgl. Abb. 7).

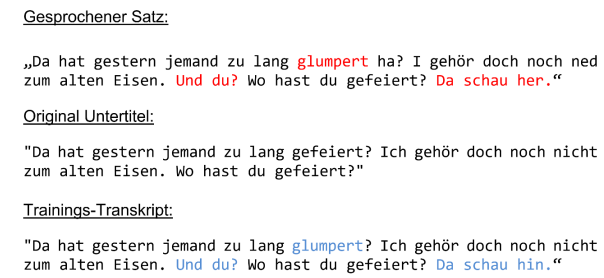

Abbildung 7: Unterschied zwischen gesprochenem Satz, original Untertitel und Trainings-Transkript

Für das Projekt „Oachkatzl“ wurden im ersten Schritt 50 Stunden bayerisches Trainingsmaterial generiert.

Training des Systems

Da für ein vollständiges Training eines AM-Systems 50 Stunden Material nicht ausreichen, und um die Qualität der Daten bzw. das Potenzial des Systems abschätzen zu können, wird hier ein Nachtraining angewendet. Dafür wird das neuronale Netz zuerst mit 550 Stunden hochdeutschem Trainingsmaterial vortrainiert. Das Material stammt unter anderem aus frei verfügbaren Trainingsdatenbanken und internen Beständen von „ainblick“. Anschließend wird das System mit den 50 Stunden bayerischen Materials nachtraininert. So lernt das System zuerst die Grundzüge deutscher Aussprache bzw. den Wortschatz. Über das bayerische Trainingsmaterial werden die Besonderheiten des bayerischen Dialekts vermittelt.

Beim Training eines neuronalen Netzes müssen unterschiedliche technische Parameter berücksichtigt werden, welche für den Trainingserfolg eine zentrale Rolle spielen. Für Training eines neuronalen Netzes auf Dialekte existieren aktuell kaum Erfahrungswerte. Deshalb müssen die optimalen Parameter evolutionär, durch das Training mehrerer neuronaler Netze mit unterschiedlichen Parameterwerten bestimmt werden.

Benchmark

Um das Potenzial eines BR-eigenen bayerischen AM-Tools und damit die Qualität der Trainingsdaten bewerten zu können, muss ein System auf die Anwendungsfälle des BRs getestet werden. Dazu dient der hier beschriebene Benchmark.

Verglichen werden das mit den BR-Daten erstellte „Oachkatzl“-Modell und das AM-Tool „SAM“ des Fraunhofer-Instituts IAIS. „SAM“ wird aktuell sowohl beim BR als auch ARD-weit produktiv eingesetzt und dient damit als Referenz-Tool.

Das „Oachkatzl“-Modell befindet sich aktuell in einem Prototypen-Status. Das bedeutet, dass dieses Modell, im Gegensatz zu SAM, kein zusätzliches Tool zur Korrektur von Groß- und Kleinschreibung oder Grammatik besitzt. Deshalb wird im Benchmark ausschließlich die Wort-für-Wort-Transkription verglichen.

Für den Benchmark werden Audio-Chunks vorbereitet, welche die Anwendungsfälle beim BR möglichst weit abdecken. Anschließend werden die Chunks, getrennt nach festgelegten Test-Kriterien, durch die AM-Tools SAM und „Oachkatzl“ transkribiert. Die automatisch erzeugten Transkripte werden dann mit vorab erstellten Chunk-Musterlösungen über die WER verglichen.

Die Berechnung der WER findet über ein speziell für diesen Benchmark angepasstes Software-Tool statt (Abrufbar unter: https://colab.research.google.com/drive/1Fqa21UjMZp1MbRcznPFsZavDJBB1BDbg).

Testkriterien

Die Qualität der Systeme kann hinsichtlich einer vorab definierten Dialektstärke bewertet werden. Die Dialektstärke beschreibt, wie stark gefärbt eine Person Dialekt spricht. Ein gutes System sollte in der Lage sein, unabhängig von der Dialektstärke konstante Analyseergebnisse zu liefern. Für den Benchmark werden die Daten in vier Kategorien eingeteilt:

Hochdeutsch (HD):

Hochdeutsch dient als Referenzmenge für die verschiedenen Sprachmodelle. Die Kategorie Hochdeutsch bezieht sich hier auf hochdeutsche, natürliche Sprache ohne dialektale Färbung. Falls ein System bereits bei hochdeutscher Sprache schlechtere Ergebnisse liefert als andere, wird dieses auch bei bayerischer Sprache schlecht abschneiden. Durch den Vergleich mit Hochdeutsch können trotzdem Systeme, die auf unterschiedlichen Qualitätsniveaus arbeiten (z.B. ein Prototyp und ein eingekauftes Tool), miteinander verglichen werden.

Medien-Bayerisch (MB):

Ein großer Anteil an bayerischer Sprache in den BR-Sendungen stammt von Moderatoren, Schauspielern oder Kabarettisten. Diese Personen können der Gruppe der professionellen Sprecher*innen zugeordnet werden. Das bedeutet, dass sie durch ihren Beruf sehr geübt im Sprechen sind und oft zusätzlich ein Sprechertraining absolviert haben. Die Verständlichkeit des Sprechers steht im Vordergrund. Dialekt dient als

Stilmittel. Dementsprechend besitzt das Medien-Bayerisch eine besonders verständliche Artikulation, und die Grammatik ähnelt stark der Schrift-Grammatik.

Gemäßigtes Bayerisch (GB):

GB stammt von nicht professionellen Sprechern zum Beispiel interviewten Personen auf der Straße. Diese Personen haben in der Regel eine undeutlichere Aussprache als professionelle Sprecher und die gesprochene Grammatik unterscheidet sich stärker von der Schrift-Grammatik (Wortwiederholungen, fehlende Artikel etc.). Außerdem stammen alle gesprochenen Wörter vom Wortstamm eines im hochdeutschen existierenden Wortes ab (z.B. „kimmst“ von „kommen“).

Schweres Bayerisch (SB):

Größtes Unterscheidungsmerkmal zwischen SB und GB ist, dass im SB dialekt-spezifische Begriffe vorkommen (z. B. „glumpert“ hochdeutsch: „gefeiert“). Außerdem ist die Artikulation für den bayerischen Dialekt oft weicher/verwaschener, und die Grammatik besteht zu großen Teilen aus dialekt-spezifischer Grammatik (z.B. doppelte Verneinung).

Ergebnisse

Bei der Bestimmung der WER für die einzelnen Test-Kriterien für das „Oachkatzl“-Modell und SAM ergeben sich folgende Werte:

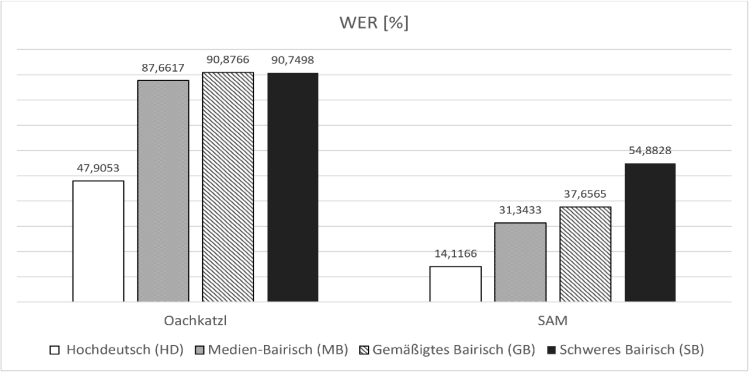

Diagramm 1: WER bei unterschiedlichen Dialektstärken für das „Oachkatzl“-Modell und SAM

Die WER des „Oachkatzl“-Modells ist bei allen Test-Kriterien wesentlich höher als bei SAM (vgl. Diagr. 1). Bezieht man mit ein, dass das „Oachkatzl“-Modell vorrangig von einer einzelnen Person der Firma „ainblick“ trainiert wurde und weniger als 600 Stunden Trainingsmaterial verwendet wurden, sind die Ergebnisse des Modells durchaus vielversprechend. Besonders die annähernd gleiche Analysequalität des „Oachkatzl“-Modells bei MB, GB und SB bei steigender Dialektstärke, unterscheidet sich deutlich von SAM (vgl. Diagr. 1). SAM liefert bei zunehmender Dialektstärke deutlich schlechtere WER-Werte. Ziel ist es ein AM-System zu entwickeln, dass bei verschiedenen Dialektstärken ähnliche Analyseergebnisse aufweist. Das Training des „Oachkatzl“-Modells im kleinen Rahmen hat gezeigt, dass dies grundsätzlich möglich ist. Für einen produktiven Einsatz müsste nun, durch das Training mit deutlich mehr Material, die WER auf einen Wert um 15% gesenkt werden.

Zusammenfassung

Zentrales Ziel dieses Projekts ist die Beantwortung der Frage, ob und wie eine Datenbank für das Training eines produktiv einsetzbaren AM-Systems für bayerische Dialekte aufgebaut werden kann. Die Trainingsdatenbank soll auf BR-eigenen Fernsehsendungen mit Untertiteln basieren.

Die Durchführung des Projekts „Oachkatzl“ mit den Ergebnissen des Benchmarks beweist, dass BR-eigene Daten grundsätzlich für ein AM-Training nutzbar sind.

Die Gegenüberstellung des „Oachkatzl“-Models und SAM hat gezeigt, dass das Training mit Fernsehsendungen und Untertiteln ein robustes AM-System, gegenüber verschiedenen Dialektstärken, hervorbringen kann. Entscheidend ist, dass eine manuelle Bearbeitung der Daten erfolgt (Korrektur der Untertitel), um ein erfolgreiches Training zu gewährleisten. Aus diesem Grund ist die Optimierung des teilautomatisierten Workflows zur Trainingsdatengenerierung besonders wichtig. Nur so ist die Generierung einer großen Trainingsdatenbank mit akzeptablem Ressourcenaufwand möglich.

Diese Arbeit hat das große Wertschöpfungspotenzial des BR-Archivs als Quelle für AM-Trainingsdaten deutlich gemacht. Darüber hinaus wurden die Methodik, die Test-Daten und das Benchmark-Tool vorgestellt, um die Qualität verschiedener AM-Systeme effektiv miteinander vergleichen zu können. Damit stehen dem BR nun die nötigen Werkzeuge zur Verfügung, um zukünftig aufgebaute Trainingsdatenbanken präzise zu bewerten.

Dieses Projekt hat damit den ersten Schritt in Richtung der Entwicklung produktiv einsetzbarer Audio-Mining-Systeme für bayerische Dialekte getan.

Prozesse des BRs welche eine Transkription von bayerischer Sprache in hochdeutschen Text benötigen, könnten dann durch diese Tools maßgeblich beschleunigt bzw. erweitert werden.